Models of memory interference in multiprocessors

L. Boguslavski, A. Greenberg, Philippe Jacquet, C.P. Kruskal, A. Stolyar

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1469

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

MODELS OF MEMORY INTERFERENCE IN MULTIPROCESSORS

Leonid BOGUSLAVSKI
Albert G. GREENBERG
Philippe JACQUET
Clyde P. KRUSKAL
Alexander STOLYAR

Juin 1991

RR-1469

Models of Memory Interference in Multiprocessors

Leonid Boguslavski, Albert G. Greenberg, Philippe Jacquet,
Clyde P. Kruskal, Alexander Stolyar

Abstract: A basic, widely used model of memory interference in multiprocessors is considered, and generalized in several directions. In part 1 a simple approach to the analysis of these systems is presented, which yields estimates of transient and equilibrium performance statistics. Simulation data show the estimates are very accurate. Part 2 gives the asymptotics of the basic model, and establishes that the estimates given here become exact as the system becomes large.

Modélisation des conflits de mémoires dans les architectures à processeurs multiples

Résumé : Nous considérons un modèle bien connu des conflits de mémoires dans les architectures à processeurs multiples que nous généralisons à l'aide de nombreuses variantes. Nous présentons dans la première partie une approche simple de l'analyse de ces systèmes qui permet une estimation de leurs états transitoires et stationnaires. Nous reportons des simulations qui confirment la grande précision des estimations. La deuxième partie considère le modèle de base d'un point de vue asymptotique et montre notamment que les estimations de la première partie deviennent exactes lorsque la taille du système croît.

L. Boguslavski and A. Stolyar are with the Institute of Control Sciences, Profsoyuznaya ul. 65, Moscow GSP-312, USSR

A. G. Greenberg is with AT&T Bell Laboratories, Room 2C-119, Murray Hill, NJ 07974, USA

P. Jacquet is with INRIA, Rocquencourt, 78153 Le Chesnay cedex, France

C.P. Kruskal is with the Dept. of Computer Science, University of Maryland, College Park, MD 20742, USA

MODELS OF MEMORY INTERFERENCE IN MULTIPROCESSSORS, PART I: APPROXIMATE ANALYSIS

Leonid B. Boguslavsky[1], *Albert G. Greenberg*[2], *Philippe Jacquet*[3], *Clyde P. Kruskal*[4], *Alexander L. Stolyar*[1]

December, 1990

Abstract: A basic, widely used model of memory interference in multiprocessors is considered, and is generalized in several directions. A simple approach to the analysis of these systems is presented, which yields estimates of transient and equilibrium performance statistics. Simulation data show the estimates are very accurate. Our companion paper gives the asymptotics of the basic model, and establishes that the estimates given here become exact as the system becomes large.

1 Institute of Control Sciences, Profsoyuznaya ul.65, Moscow GSP-312, USSR
2 Room 2C-119, AT&T Bell Laboratories, Murray Hill, NJ, USA 07974
3 INRIA, Rocquencourt, 78153 Le Chesnay Cedex, France
4 Dept. of Computer Science, University of Maryland, College Park, MD 20742 USA

1. INTRODUCTION

One method of building a parallel computer is to connect N processors to M memory modules via a crossbar switch, as depicted in Figure 1*a*. In a very simple model of system behavior, which we call the *basic model*, time is divided into discrete cycles. At the start of each cycle, some of the processors will have memory requests pending, and the remaining processors will not. Each processor with a request pending to a given memory module reissues its request to the same module. Each processor with no request pending accesses a memory module chosen independently at random, with each choice equally likely. Next, each memory services one of the requests targeted to it (if there is one). More generally, one can view this as a queueing system, as depicted in Figure 1*b*. Rather than having a processor continually reissue a request, requests are queued at the memory modules. At the beginning of a cycle, each processor with no queued request issues a new request to some random memory module. Next, each memory module with a nonempty queue, services a request.

This basic model of a parallel computer can be generalized in various ways. Processors could compute (think) for a random amount of time before issuing a request. Requests can take a random amount of time to service. A processor could have a preferred module or set of modules (spacial locality). A processor could be more likely to send a request to the module that it just sent a request to (temporal locality). A processor could issue several requests at the same time (e.g., for a set of memory locations comprising a cache line).

A fundamental problem is to analyze the performance of such systems. There are many questions one might ask, including: What is the bandwidth of the memory system? How long does it take a request to be satisfied? What is the distribution of the queue lengths? What is the transient behavior, e.g. how does the system behave from start up? Previous analyses have only considered steady state behavior. There have been two approaches. One is to obtain exact analyses [2,13,15], which has been successful only for small M or N. The other approach is to approximate the behavior [1,2,8,12,16,18,19,20]. This has lead to good approximations, but there has been no proof of correctness (i.e., no rigorous connection has been established between the approximations and the actual model).

In this paper, we present a general queueing theoretic analysis of the model. Previous queueing analyses [1,19] have taken a "local" view of the problem; we take a "global" view, which has several advantages. First, we obtain more general formulae. The probability distribution of the think time may be any distribution on the nonnegative integers having a finite first moment. The probability distribution of the service time at a memory module may be any distribution on the positive integers having finite first and second moments. We can allow some temporal and spacial locality. Second, we can analyze transient behavior, as well as steady state. Baskett and Smith [1] presented an approximate steady state analysis of the basic model. While it is generally believed that their approximation is "asymptotically" accurate [9,11,16,20] there has been no proof of this fact. Our "global" view allows us -- in a companion paper [5] -- to show that their approximation, generalized to include geometric think times, is indeed "asymptotically" accurate.

Large parallel machines cannot afford the cost of a crossbar switch. Instead, machines often use a multistage interconnection network, such as an Omega network (see [14] for a survey of interconnection networks). The purpose of this paper is to try to understand how memory conflicts affect performance, without the complications of an interconnection network. In a future paper, we will show how our analyses can be extended to incorporate an interconnection network.

Subsection 1.1 surveys previous work. Section 2 analyzes the basic model described initially, except processors think for a geometrically distributed amount of time. We call this the basic geometric think time model. Section 3 presents a series of more general models:

1. **General Think Times:** The probability distribution of the think time may be any distribution on the nonnegative integers having a finite first moment.

2. **General Service Times:** The probability distribution of the time a memory needs to service a request may be any distribution on the positive integers having finite first and second moments.

3. **Access Correlations:** In practice, there may be correlation between accesses. To model this, we suppose the memories are partitioned into m classes. When a processor stops thinking it chooses which memory class to access next, and then accesses a memory node in that class chosen uniformly

at random. A Markov chain governs the choice of class: A processor chooses class j with probability $\xi_{i,j}$ where i is the class last accessed; $1 \le i,j \le m$. We suppose that the $\xi_{i,j}$ determine an aperiodic and irreducible Markov chain.

4. **Fork/Join Accesses:** In practice, a processor might issue a request for a block of memory registers, potentially held in different memory nodes. (An example of this is a cache line, which is typically a set of contiguous memory locations.) To model this, we suppose that when a processor stops thinking it splits into c requests (the fork), each of which then queues at a memory node chosen independently and uniformly at random, where $c \ge 1$ is a parameter. After all c of the requests have received service at the memory nodes, the processor returns to thinking.

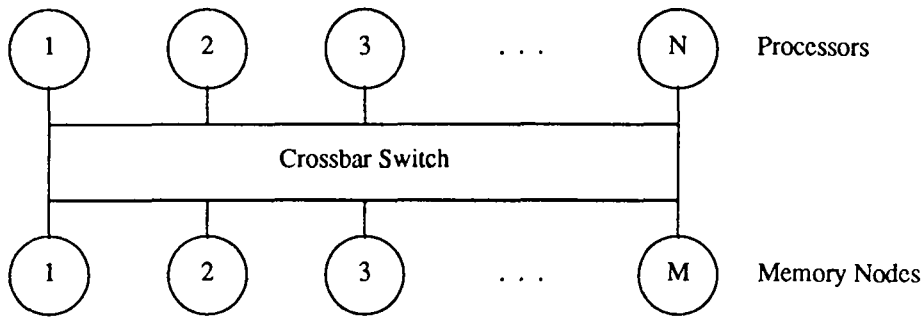

Figure 1a. Multiprocessor model.

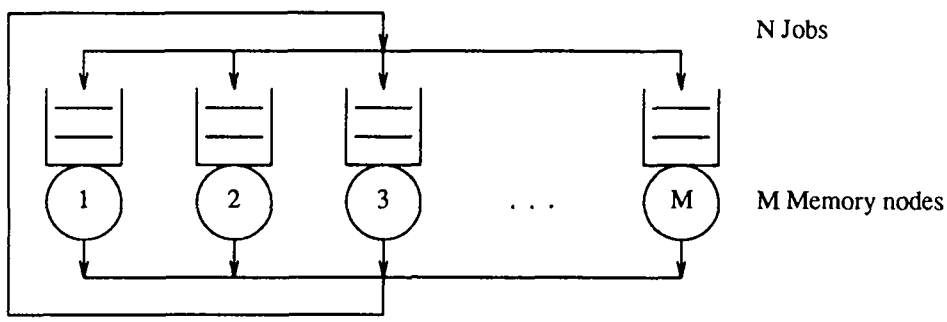

Figure 1b. Multiprocessor model viewed as a queueing network.

1.1 PREVIOUS WORK

Previous work has centered on finding the steady-state utilization of a multiprocessor memory system. See Yen et al. [20] for a survey. No useful closed form solutions have been obtained for general M and N.

For (small) fixed M and N researchers have used Markov chains to find exact solutions. Skinner and Asher [15] studied the basic model (no think time and constant service time), and obtained exact results for $N=2$. Bhandarkar [2] obtained exact results for small M and N. Sethi and Deo [13] studied what happens if a processor is more likely to send a request to the same module to which it just sent one (temporal locality), and obtained exact results for small N.

There have been two approaches to approximating the performance for general M and N. The first, originally used by Strecker [18], makes the simplifying assumption that requests that have not been serviced at a particular cycle are resubmitted to a random memory module, rather than being queued at the same memory module. This means that the requests are independent, which simplifies the analysis. Rau [12] presented a summation formula for the basic model, which turns out to be equivalent to Strecker's (closed form) approximation in this special case. Bhandarkar [2] improved Strecker's formula for the basic model by noticing that the memory utilization is almost symmetric in M and N. Hoogendoorn [8] considers the situation where each processor accesses the memory modules with a general probability distribution. Smilauer [16] uses "mean value analysis" to improve on Hoogendoorn's results.

Baskett and Smith [1] presented an approximate steady state analysis of the basic model, which provides, in particular, the average queue length and the utilization of any given memory node. They generalize this in an ad hoc way to approximate geometric think time (and constant service time). Yen and Fu [19] generalize Baskett and Smith's work in a more natural way to obtain a better approximation for geometric think time (and constant service time).

In brief, Baskett and Smith's approach for the basic model was to focus on the queue at a single memory node and model that essentially as an M/D/1 queue [10] receiving requests at an unknown rate λ. The parameter λ was then determined by solving the M/D/1 model for the average queue length and insisting this average equal N/M. Our approach is different, though the equilibrium results are essentially

the same for the model of Figure 1*b*. We work with the state of the entire system, describing for every $i \geq 0$ the proportion of memory node queues that hold i requests. We then derive an "expected move" operator that maps the current state into an expected next state, as described below. This leads to approximations for the transient (finite time) behavior of the system. The unique fixed point of the expected move operator provides approximations for the equilibrium behavior of the system.

Fukuda uses "equilibrium point analysis" to approximate memory systems with combinations of geometric and zero think times and constant and geometric service times. Each case is analyzed separately and (unlike previous work) the results only take into account the ratio of M and N, but not their direct values. Fukuda also considers access correlations and dynamic reference patterns, but his solutions depend on solving systems of nonlinear equations.

2. BASIC GEOMETRIC THINK TIME MODEL

In this section we present our *basic geometric think time model* of a multiprocessor and its approximate transient and steady state analysis. Throughout, time is discrete and is counted in *cycles*, where the t-th cycle spans the interval $[t,t+1)$. The number of processors is N and the number of memory nodes is M. At the start of a cycle, some processors will have a memory request pending (queued for service), and some will not. Those with no request pending are said to be *thinking*. A cycle consists of two consecutive actions:

1. Independently with probability p, each thinking processor stops thinking and requests access to one of the M memory nodes, chosen uniformly at random. With probability $1-p$, the processor remains thinking.

2. Each memory unit serves exactly one request (provided it has one), and the associated processor returns to thinking.

Hence, a memory unit has no request pending at time $t+1$ if it had none at time t and received zero or one request in cycle $[t,t+1)$, or the unit had one at time t and received none at cycle $[t,t+1)$.

For purposes of analysis, the system is best represented as the queueing network depicted in Figure 2. There are N jobs, corresponding to the N processors. A job holds at the think node for a geometrically

distributed number of cycles with mean $1/p - 1$, and then queues at one of the *M memory nodes*, chosen uniformly at random. A memory node serves one job per cycle (if it has at least one), returning that job to the think node. We refer to the holding time at the think node as the *think time*. (A job incurs a think time of 0 if it leaves for a memory node at the next cycle after returning from a memory node.)

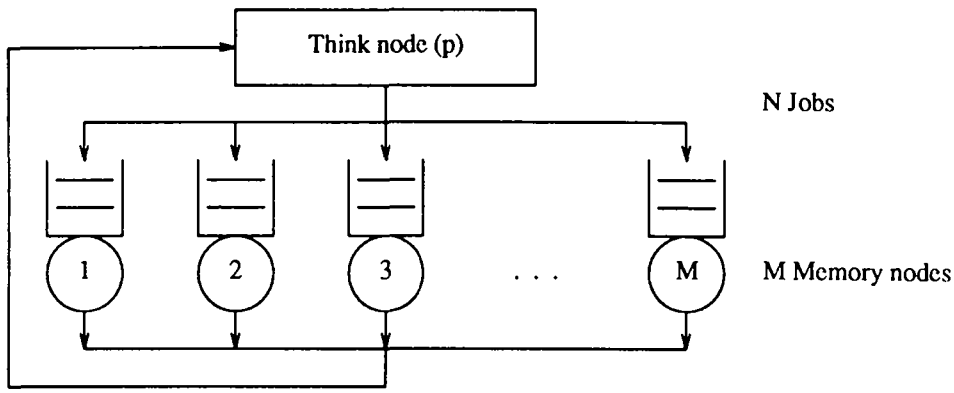

Figure 2.
Basic geometric think time model.

Since any of the N jobs can wait in any of the M queues, the number of distinct configurations is at least M^N. The key to the analysis is to avoid this huge and unwieldy state space, and deal with the smaller space where a state tells only how many queues hold 0, how many hold 1, how many hold 2, and so forth. The sequence of consecutive states is a Markov chain. A simple operator describes how the average values of these quantities change in time, conditioned on their values at time $t = 0$. We obtain sharp estimates of the transient behavior of the system at small computational cost from that operator. The operator has a unique fixed point, which is easy to obtain explicitly. The fixed point provides sharp estimates of steady state behavior.

In the analysis, we use generating functions – in a somewhat unusual way – to encode the state of the system. These *state generating functions* are random and should not be confused with probability generating functions, though we will see that the two are related. In this section, the generating functions are helpful in streamlining the presentation. In the companion paper [5] on the asymptotics of the

approximations, the generating functions are essential.

To describe the state of the system, let q_i^t denote the proportion of memory nodes that hold i jobs at time t, that is,

$$M\, q_i^t = \text{ the number of memory nodes that hold } i \text{ jobs at time } t\ .$$

Define

$$q^t(z) = \sum_{i=0}^{N} q_i^t\, z^i\ ,$$

a generating function that represents the *state* at time t. Let

$$\Omega_{N,M}$$

denote the set of reachable states. We see that time t a total of

$$M\,(q^t)'(1) = M\,\frac{dq^t}{dz}(1) = M\sum_i i\, q_i^t$$

jobs reside at memory, and a total of

$$N - M(q^t)'(1) = M(\alpha - (q^t)'(1))$$

reside at the think node, where

$$\alpha = \frac{N}{M}$$

is the ratio of processors to memory nodes. For every state $q \in \Omega_{N,M}$,

$$q'(1) \le \alpha\ ,$$

since the maximum number of jobs distributed among the M memories is N.

The connection between these state generating functions and probability generating functions stems from

$$\mathrm{E}\,(q_i^t) = \Pr\,(\text{ a given memory node holds } i \text{ jobs at time } t\ )\ ,$$

which holds by symmetry. Formally define

$$E\, q^t(z) = \sum_i E\,(q_i^t)\, z^i \ ,$$

and note that this is the probability generating function for the queue length at a given memory node at time t.

APPROXIMATIONS

Our approximation is to pretend that q^t always takes the expected move, characterized by the operator

$$\phi(q) = E(q^{t+1} \mid q^t = q) \ .$$

Specifically, the approximation is

$$q^t = E(q^t) = \phi(q^{t-1}) = \phi^t(q^0) \ , \text{ where}$$

$$\phi^0(q) = q \ \ ; \ \ \phi^t(q) = \phi(\phi^{t-1}(q)) \quad \text{for } t \geq 1 \ .$$

We interpret $\phi(q^t)$ as an estimate of q^{t+1}, the real system state. Our approximation for q^t at steady state is

$$q^* = \phi(q^*)$$

the fixed point of ϕ, which is given below, and is unique among the positive generating functions $q(z)$ with $q(1) = 1$ and $q'(1) \leq \alpha$.

We can routinely extract performance estimates from these approximations. Suppose q^0 is given and let

$$r^t = \phi^t(q^0) \ .$$

Take

$$r_i^t \ \ ; \ i = 0,1,2,... \ , \ t = 0,1,2,...$$

as an estimate of the proportion of memory nodes that hold i jobs at time t, and

$$(r^t)'(1) \ \ ; \ t = 0,1,2,...$$

as the average number of jobs in the queue of a given memory node at time t. Higher order derivatives of $r^t(z)$ provide estimates of higher order moments of the queue length. Taking $M(r^t)'(1)$ as an estimate of the average number of jobs at memory, we are led to the following estimate of the average number of jobs

that arrive to a given memory node in cycle $[t,t+1)$,

$$\lambda^{t+1} = p\,(\alpha - (r^t)'(1)) \;\; ; \;\; t = 0,1,2,....$$

In addition, we obtain the following estimate of the utilization of a given memory node in cycle $[t,t+1)$, that is, the probability that a memory node serves a job in that cycle,

$$u^{t+1} = 1 - r_0^t\,(1 - \frac{p}{M})^{M(\alpha - (r^t)'(1))} \; . \tag{2.1}$$

To obtain corresponding steady state estimates, we just replace r^t with the fixed point of the expected move operator, $q^* = \phi(q^*)$. In so doing we obtain $\lambda^* = u^*$, where λ^* and u^* are the counterparts of λ^{t+1} and u^{t+1}. This is to be expected since at equilibrium jobs depart from the think node at the same rate as they depart from the memory nodes. Closed forms for q^* and λ^* are given in Proposition 2.1 below.

Next, consider the equilibrium response time. Assuming, for example, that the queuing disciple is FIFO, an estimate of the equilibrium response time distribution can be derived from the results to follow by standard methods. By Little's law, at equilibrium the average response time is the ratio of the queue length at a given memory node to the node's throughput. Our estimate of the equilibrium response time is therefore

$$1 + \frac{(q^*)'(1)}{\lambda^*} \; .$$

We have the following simple expressions for ϕ and the estimates derived from ϕ:

Proposition 2.1: *Our estimate of the utilization of a memory node at equilibrium is* λ^*, *the unique root* $\lambda = \lambda^*$ *in* $[0,1]$ *of the quadratic*

$$\alpha = \frac{\lambda}{p} + \frac{\lambda\,(\lambda - p/M)}{2(1 - \lambda)} \; , \tag{2.2}$$

namely

$$\lambda^* = \frac{\alpha p + 1 - p^2/(2M) - \sqrt{(\alpha p + 1 - p^2/(2M))^2 - 2\alpha p(2 - p)}}{2 - p}$$

where $\alpha = N/M$. *The expected move operator,* $\phi(q) = \mathrm{E}(q^{t+1} \mid q^t = q)$, *is given by*

$$\phi(q(z)) = \frac{q(z)\ a(z) - q(0)\ a(0)}{z} + q(0)\ a(0) \tag{2.3a}$$

$$a(z) = \left[1 + \frac{p}{M}(z-1)\right]^{M(\alpha - q'(1))} . \tag{2.3b}$$

Our estimate of the system state at equilibrium is the fixed point $\phi(q) = q$, *which is given by*

$$q^*(z) = \frac{(1 - \lambda^*)\ (z - 1)}{z - (1 + \frac{p}{M}(z-1))^{M\lambda^* / p}} . \tag{2.4}$$

It is easy to compute $\phi^t(q^0)$ quickly; iteratively apply ϕ as given in (2.3a), truncating the generating functions where the mass lost is negligible. With $\phi^t(q^0)$ in hand, all other statistics of the transient analysis are simple to compute. All statistics of the steady state analysis are trivial to compute.

The analysis leading to the equilibrium estimates (2.3a) and (2.3b) is very similar to the analysis of the M/D/1 queue [10], and is similar to Baskett and Smith's original treatment [1]. Fix q in $\Omega_{N,M}$, drop superscripts and let

$$q^t(z) = q(z) = \sum_{i=0}^{\infty} q_i\ z^i$$

$$\phi(q(z)) = r(z) = \sum_{i=0}^{\infty} r_i\ z^i .$$

In cycle $[t,t+1)$, i jobs leave the think node and arrive to a given memory node with the binomial probability,

$$a_i = \binom{N - Mq'(1)}{i} \left[\frac{p}{M}\right]^i \left[1 - \frac{p}{M}\right]^{N - Mq'(1) - i} .$$

The corresponding probability generating function is

$$a(z) = \sum_{i=0}^{\infty} a_i\ z^i = \left[1 + \frac{p}{M}(z-1)\right]^{N - Mq'(1)} , \tag{2.5}$$

which is equivalent to (2.3b). The quantity Mr_i represents the expected number of memory nodes that will hold i jobs at time $t+1$, $i \geq 0$. If $i \geq 1$, then such a node must hold j jobs at time t, for $0 \leq j \leq i+1$, and must receive $i+1-j$ new jobs in cycle $[t,t+1)$. Similarly, if $i = 0$, the node must either hold no jobs at

time t and receive zero or one in cycle $[t,t+1)$, or hold one job at time t and receive none in cycle $[t,t+1)$. Since Mq_i^t is the number of nodes that hold i jobs at time t,

$$r_0 = q_0\, a_0 + q_0\, a_1 + q_1\, a_0$$

$$r_i = q_0 a_{i+1} + q_1 a_i + \cdots + q_{i+1} a_0 \quad ; \quad i \geq 1 \ .$$

This sytem is easily manipulated to give (2.3a).

Next, to find the fixed point of the expected move operator ϕ, suppose

$$q(z) = \phi(q(z)) = \frac{q(z)a(z) - q(0)a(0)}{z} + q(0)a(0) \ ,$$

so

$$q(z) = q(0)a(0)\ \frac{z-1}{z-a(z)} \ .$$

Let

$$\lambda = a'(1) = p(\alpha - q'(1)) \ ,$$

and note

$$a''(1) = \frac{p^2}{M^2}(N - Mq'(1))\ (N - Mq'(1) - 1) = \lambda\ (\lambda - \frac{p}{M}) \ .$$

Applying L'Hopital's rule to evaluate $q(1)$ gives

$$q(1) = \frac{q(0)a(0)}{1 - a'(1)} = \frac{q(0)a(0)}{1-\lambda} \ ,$$

which, since $q(1) = 1$, implies $q(0)a(0) = 1 - \lambda$, and

$$q(z) = (1-\lambda)\ \frac{z-1}{z-a(z)} \ . \qquad (2.6)$$

Again, by L'Hopital's rule,

$$q'(1) = \frac{1}{2}\ \frac{a''(1)}{(1-\lambda)} = \frac{\lambda\ (\lambda - p/M)}{2(1-\lambda)} \ .$$

But by the definition of λ,

$$q'(1) = \alpha - \frac{\lambda}{p} \tag{2.7}$$

Equating the two expressions for $q'(1)$ gives (2.2), determining $\lambda = \lambda^* \in [0,1]$. The expression for the fixed point, (2.4), follows from (2.7) and the expressions for $a(z)$ and $q(z)$, (2.5) and (2.6), respectively.

The formulae of Proposition 2.1 all simplify in the limit

$$\text{as } N \to \infty \text{ with } \frac{N}{M} \to \alpha$$

where $\alpha > 0$ is an arbitrary constant. Specifically, for every z, $\phi(q(z))$ converges to

$$\sigma(q(z)) = \frac{q(z)a(z) - q(0)a(0)}{z} + q(0)a(0) \tag{2.8a}$$

$$a(z) = e^{\lambda(z-1)} \quad , \quad \lambda = p(\alpha - q'(1)) \tag{2.8b}$$

since $a(z)$, as given by the binomial probability generating function (2.3b), converges to the Poisson probability generating function with the same mean. Solving for the fixed point of σ leads to the same results as given in Proposition 2.1 except that the terms with $1/M$ factors vanish. These results are crucial to the theoretical justification of the approximations given in the companion paper [5]. However, they are less accurate than the formulae of Proposition 2.1 (when compared with Monte Carlo simulation results), and are no easier to compute.

We carried out a wide range of experiments, testing the accuracy of approximations derived from the iterates

$$\phi^t(q^0) \quad ; \quad t \geq 1$$

with the initial condition

$$q^0(z) = 1 \ ,$$

meaning all jobs start at the think node. This entailed running Monte Carlo simulations of the model over a finite time horizon $1, \ldots, T$, and collecting the following statistics from the iterates and the simulation:

- $(q^t)'(1)$, the mean number of jobs queued at one memory node,
- λ^{t+1}, the mean number of jobs that arrive to a memory node in cycle $[t, t+1)$, and

- u^{t+1}, the average utilization of a memory node in cycle $[t,t+1)$.

Over a wide range of parameter values, we found the relative error between the estimates and the simulation data to be uniformly less than 1%. In Figure 3 we plot the two data sets for $(q^t)'(1)$, for $t = 1, 2, ...,25$, and $N = 10, M = 5, p = 1$ (so the think time is 0). The widths of the 95% confidence intervals are less than .01% of the values plotted.

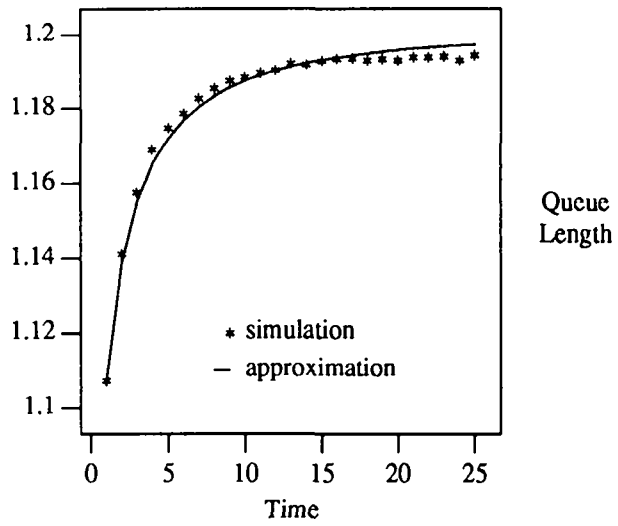

Figure 3. Queue length at a memory node versus time, for $N = 10$ processors, $M = 5$ memory nodes, and think time parameter $p = 1$ (0 think time).

The accuracy of the steady state approximations is comparable. By symmetry, at equilibrium the total throughput is the product of M and the equilibrium memory utilization utilization. Our estimate is $M\lambda^*$, where λ^* is the root in $[0,1]$ of (2.2). In Table 1, we refer to this root as the *binomial estimate*, and also consider a *Poisson estimate*, defined as the root of the simpler counterpart of (2.2) obtained by letting $M \to \infty$.

Looking along the diagonals in Table 1 , we see that, as N and M become large, doubling both N and M almost exactly doubles the throughput. The utilization (throughput per memory node) is obtained by dividing the throughput by M. The Table shows that, for large N and M, the utilization depends on N and M essentially only through the ratio of N to M. Thus, the Poisson estimate becomes sharper as N and M increase. The binomial estimate is remarkably sharp even if N and M are both small. The Table also shows

that throughput is a nearly symmetric function of N and M, as Rau noticed [1,12].

		Number of Memory Nodes, M					
		2	4	8	16	32	64
N	2	1.500 [-4.093%] (-21.88%)	1.750 [-1.418%] (-12.69%)	1.875 [-0.382%] (-6.464%)	1.938 [-0.100%] (-3.203%)	1.969 [-0.024%] (-1.584%)	1.984 [-0.006%] (-0.786%)
u m b e r	4	1.750 [-1.431%] (-12.70%)	2.621 [-1.728%] (-10.60%)	3.265 [-0.803%] (-6.406%)	3.627 [-0.250%] (-3.284%)	3.813 [-0.069%] (-1.620%)	3.906 [-0.018%] (-0.789%)
o f	8	1.875 [-0.387%] (-6.469%)	3.266 [-0.836%] (-6.437%)	4.948 [-0.859%] (-5.283%)	6.314 [-0.414%] (-3.212%)	7.134 [-0.143%] (-1.659%)	7.564 [-0.040%] (-0.815%)
P r o c e s s o r s , N	16	1.938 [-0.115%] (-3.218%)	3.627 [-0.271%] (-3.305%)	6.315 [-0.433%] (-3.230%)	9.626 [-0.421%] (-2.631%)	12.423 [-0.214%] (-1.612%)	14.147 [-0.065%] (-0.823%)
	32	1.969 [-0.039%] (-1.599%)	3.813 [-0.071%] (-1.621%)	7.133 [-0.140%] (-1.657%)	12.423 [-0.213%] (-1.611%)	18.994 [-0.213%] (-1.308%)	24.643 [-0.203%] (-0.798%)
	64	1.985 [-0.022%] (-0.803%)	3.906 [-0.013%] (-0.793%)	7.563 [-0.028%] (-0.804%)	14.146 [-0.059%] (-0.817%)	24.645 [-0.111%] (-0.810%)	37.738 [-0.104%] (-0.657%)

Table 1. Equilibrium throughput, measured in simulations, for a range of M and N, with $p = 1$ (0 think time). The numbers in square brackets give the relative errors of the binomial estimates, and those in parentheses the relative error of the Poisson estimates. The widths of the 95% confidence intervals computed for the simulation data were never greater than 0.03% of the values in the Table.

3. MORE GENERAL MODELS

In this section, we consider four generalizations of the basic geometric think time model, allowing: general think time probability distributions, general memory service time probability distributions, access correlations controlled by a Markov chain, and job fork/join. For each of the first three generalizations, the analysis is similar to the analysis of the basic geometric think time model:

1. The global state at time t is defined telling the number of jobs that belong to each substate, but not which jobs belong to each substate.

2. A tractable expression is derived for the expected move operator, which accounts the expected change in each substate in one cycle.

3. The fixed point of the expected move operator is derived. This yields equilibrium performance statistics.

We also indicate how the three generalizations can be combined. For the fork/join model, the method appears to break down because of the complexity of the global state. However, a further (apparently accurate) approximation removes that stumbling block, and allows us to derive equilibrium performance statistics.

3.1 GENERAL THINK TIME MODEL

We generalize the basic geometric think time model by allowing the think time distribution to be any distribution on the non-negative integers having a finite mean. Everything else remains the same. To describe the think time distribution, let

$$R_i = \Pr(\text{think time} \geq i) \quad ; \quad i \geq 0$$

$$S_i = R_i - R_{i+1} = \Pr(\text{think time} = i) \quad ; \quad i \geq 0 \ .$$

Thus,

$$p_i = S_i / R_i \quad ; \quad i \geq 0$$

denotes the probability that a job departs the think node after having waited there i units of time. Two functions of this distribution turn out to be important:

$$\frac{1}{p} = 1 + \sum_{i=0}^{\infty} i S_i = 1 + \mathrm{E}(\text{think time}) \tag{3.1.1}$$

$$C = \sum_{i=0}^{\infty} (S_i)^2 / R_i \ . \tag{3.1.2}$$

To define the state of the system, we group jobs at the think node by the amount of time they have held there. We say a job residing at the think node at time t is at *stage* $i \geq 0$ if the job last entered the think node in cycle $[t-i-1, t-i)$. Define $\tau^t = (\tau_0^t, \tau_1^t, \ldots)$ and $q^t(z) = \sum_{i=0}^{\infty} q_i^t \, z^i$, where

$$N\tau_i^t = \text{number of jobs at the think node at stage } i \text{ at time } t \quad ; \quad i \geq 0$$

$$Mq_i^t = \text{number of memory nodes that hold } i \text{ jobs at time } t \quad ; \quad i \geq 0 \ .$$

The system state at time t is the pair (q^t,τ^t). Let $E\tau^t = (E\tau_0^t\ ,\ E\tau_1^t,...)$ and $Eq^t(z) = \sum_{i=0}^{\alpha} Eq_i^t\ z^i$.

Proposition 3.1: *Our estimate of the utilization of a memory node at equilibrium is* λ^*, *the unique root* $\lambda = \lambda^*$ *in* $[0,1]$ *of the quadratic*

$$\alpha = \frac{\lambda}{p} + \frac{\lambda^2 - \lambda\, C/M}{2(1-\lambda)}\ , \tag{3.1.3}$$

namely,

$$\lambda^* = \frac{\alpha p + 1 - pC\,/\,(2M) - \sqrt{(\alpha p + 1 - pC\,/\,(2M))^2 - 2\alpha p(2-p)}}{2-p} \tag{3.1.4}$$

where $\alpha = N/M$, *and* p *and* C *are functions of the think time distribution given by (3.1.1) and (3.1.2). The expected move operator*

$$\phi(q,\tau) = E\left[(q^{t+1},\tau^{t+1})\ |\ (q^t,\tau^t) = (q,\tau)\right]$$

is given by,

$$Eq^{t+1}(z) = \frac{q^t(z)a^t(z) - q^t(0)a^t(0)}{z} + q^0(0)a^t(0) \tag{3.1.5a}$$

$$a^t(z) = \prod_{i=0}^{\infty}\left[1 + \frac{p_i}{M}(z-1)\right]^{N\tau_i^t} \tag{3.1.5b}$$

$$E\tau_0^{t+1} = \frac{1}{\alpha}\left[1 - q^t(0)\,a^t(0)\right] \tag{3.1.6c}$$

$$E\tau_i^{t+1} = (1-p_i)\,\tau_i^t \quad ;\quad i \geq 1\ . \tag{3.1.6d}$$

Our estimate of the system state at equilibrium is the fixed point $\phi(q,\tau) = (q,\tau)$, *which is given by*

$$q(z) = (1-\lambda^*)\frac{z-1}{z-a(z)} \quad ;\quad a(z) = \prod_{i=0}^{\infty}\left[1 + \frac{p_i}{M}(z-1)\right]^{N\tau_i}$$

$$\tau_i = R_i\lambda^*/\alpha \quad ;\quad i \geq 0\ . \tag{3.1.7}$$

To derive the expected move operator, consider the jobs that may arrive to a given memory node in cycle $[t,\ t+1)$. The think node holds $N\tau_i^t$ jobs at stage i, each of which arrives to the memory node in question independently with probability p_i/M. Thus, each stage contributes a binomially distributed

number of arrivals, leading to the generating function (3.1.5b). By the same reasoning as used for the basic geometric think time model, this establishes (3.1.5a). Among the jobs at memory at time t, on average $M(1-q'(0)a'(0))$ are utilized in cycle $[t,t+1)$, and so return jobs to the think node in this cycle. This establishes (3.1.6c). Finally, (3.1.6d) follows since a job at stage i at time t goes to stage $i+1$ at time $t+1$ with probability $1-p_i$.

It remains to derive the fixed point of the expected move operator ϕ. Suppose

$$\mathrm{E}q_i^{t+1} = q_i^t = q_i \quad ; \quad i \geq 0$$

$$\mathrm{E}\tau_i^{t+1} = \tau_i^t = \tau_i \quad ; \quad i \geq 0 .$$

First note

$$\tau_{i+1} = (1-p_i)\tau_i = (1-p_i)\cdots(1-p_0)\tau_0 = R_{i+1}\tau_0 \tag{3.1.8}$$

and, letting $\lambda = a'(1)$,

$$\lambda = \alpha\sum_{i=0}^{\infty} p_i\tau_i = \alpha\tau_0\sum_{i=0}^{\infty} p_iR_i = \alpha\tau_0 . \tag{3.1.9}$$

Equations (3.1.8) and (3.1.9) establish (3.1.7), with $\lambda^* = \lambda$ as yet undetermined. Again, reasoning as we did for the basic geometric think time model,

$$q(z) = (1-\lambda)\frac{z-1}{z-a(z)}$$

leading to

$$q'(1) = \frac{\lambda^2 - \lambda C/M}{2(1-\lambda)} . \tag{3.1.10}$$

Accounting for the placements of all N jobs gives

$$N = N\sum_{i=0}^{\infty}\tau_i + Mq'(1)$$

which together with $\tau_i = R_i\lambda/\alpha$ leads to

$$\alpha = \lambda/p + q'(1) . \tag{3.1.11}$$

Combining (3.1.10) and (3.1.11) gives (3.1.3), determining $\lambda = \lambda^*$.

Several remarks are in order. The memory node utilization λ^* depends on the distribution of the think time only through the parameters p and C. If the think time is geometrically distributed, then $C = p$, and (3.1.3) agrees with (2.2). If, for example, the think time is of constant duration then $C = 1$. It is apparent from (3.1.4) that as N and M increase, C contributes less and less. In the limit, as $N, M \to \infty$ with $N/M \to \alpha$, all queue length statistics are insensitive to the shape of the distribution except through the mean $1/p$. In this limit,

$$a^t(z) = e^{\lambda(z-1)} \; ; \; \lambda = \alpha \sum_{i=0}^{\infty} p_i \tau_i^t \tag{3.1.12}$$

and the expected move operator ϕ tends to the operator differing from (3.1.6) only in that (3.1.12) replaces (3.1.5b). Moreover, in this limit, the expected move operator (3.1.6) coincides with the expected move operator (2.8) for the basic geometric think time model.

In Table 2 we give the results of simulations, which test (i) the sensitivity to the shape of the think time distribution, and (ii) the accuracy of the approximations. A distribution whose variance is easily controlled through a single parameter L was used: With probability $1 - 1/L$ the think time is 0 and with probability $1/L$ the think time is L. Thus, the mean is identically 1, and the variance $L - 1$. The parameter C is $1 - 1/L + 1/L^2$. The data show that as the variance becomes large, throughput remains relatively constant. The throughput estimate (which, of course, does not involve any higher moments of the think time distribution) is quite good (the relative error is less than 1%). As the variance becomes large, the error in the estimate remains relatively constant.

Think Time Parameter, L					
2	4	8	16	32	64
6.856 [−0.001%]	6.848 [+0.225%]	6.845 [+0.416%]	6.844 [+0.529%]	6.841 [+0.628%]	6.839 [+0.676%]

Think Time Parameter, L					
128	256	512	1024	2048	4096
6.840 [+0.683%]	6.836 [+0.742%]	6.841 [+0.681%]	6.845 [+0.621%]	6.860 [+0.400%]	6.862 [+0.376%]

Table 2. Equilibrium throughput, measured in simulations, with think times chosen to be 0 with probability $1-1/L$ and L with probability $1/L$, and $N=M=16$. The numbers in square brackets give the relative errors of the analytic estimate. The widths of the 95% confidence intervals computed for the simulation data were never greater than 0.6% of the values in the Table.

3.2 GENERAL SERVICE TIME MODEL

Now, let us consider the model that differs from the basic geometric think time model only in that the memory service time distribution is allowed to be any distribution on the positive integers having finite first and second moments. (In the basic geometric think time model the service time is 1 with probability 1.) To describe that distribution, let

$$r_i = \Pr(\text{service time} \geq i) \quad ; \quad i \geq 1$$

$$s_i = r_i - r_{i+1} = \Pr(\text{service time} = i) \quad ; \quad i \geq 1$$

so

$$p_i = s_i / r_i$$

is the probability that a service period that reaches i units of time stops after the i-th unit, and

$$\beta(z) = \sum_{i=1}^{\infty} s_i z^i$$

is the probability generating function for the service time.

Though our results are insensitive to the queueing disciple, for concreteness suppose that the memory nodes serve jobs FIFO. The *stage* index of a memory node is 1 plus the number of cycles of service the job in service has accumulated. Thus, a memory node at stage $i \geq 1$ at time t completes a service in cycle

$[t,t+1)$ with probability p_i.

To describe the system state, first we account for the memory nodes at stage i. For $i \geq 1, j \geq 0$, let

$$Mq_j^{t,i} = \text{number of memory nodes whose server is at stage } i \text{ with } j \text{ other jobs in queue at time } t$$

$$q^{t,i}(z) = \sum_{j=0}^{\infty} q_j^{t,i} z^j \ .$$

Let

$$q^{t,0} = 1 - \sum_{i=0}^{\infty} q^{t,i}(1) \ ,$$

so the number of idle nodes at time t is $Mq^{t,0}$, and the number of jobs at the think node is $N\tau^t$ where

$$\tau^t = 1 - \frac{1}{\alpha} \sum_{i=1}^{\infty} \left[(q^{t,i})'(1) + q^{t,i}(1) \right] \ , \tag{3.2.1}$$

since the second term in the sum accounts for the number of jobs in service at memory and the first term for the number of other jobs at memory. Let

$$q^t = (q^{t,0}, q^{t,1}, q^{t,2}, \ldots) \ .$$

The system state at time t is the pair (q^t, τ^t). As before, define the expected state to be $Eq^t = (Eq^{t,0}, Eq^{t,1}, Eq^{t,2}, \ldots)$, where $Eq^{t,i}(z) = \sum_{j \geq 0} Eq_j^{t,i} z^j$, for $i \geq 1$.

Proposition 3.2: *Our estimate of the utilization of a memory node at equilibrium is* λ^*, *the unique root* $\lambda = \lambda^*$ *in* $[0,1]$ *of the quadratic*

$$\alpha = \frac{\lambda}{p} + \frac{\mu}{2} \frac{\lambda(\lambda - p/M)}{1-\lambda\mu} + \frac{\lambda^2}{2} \frac{\mu_2 - \mu}{1-\lambda\mu} + \lambda(\mu - 1) \tag{3.2.2}$$

where $\alpha = N/M$, *and* μ *and* μ_2 *are the first two moments of the service time distribution. The expected move operator*

$$\phi(q,\tau) = E\left[(q^{t+1}, \tau^{t+1}) \mid (q^t, \tau^t) = (q,\tau) \right]$$

is given by,

$$Eq^{t+1,i}(z) = (1-p_{i-1})q^{t,i-1}(z)a^t(z) \quad ; \quad i > 2 \tag{3.2.3a}$$

$$Eq^{t+1,2}(z) = (1-p_1)q^{t,1}(z)a^t(z) + q^{t,0}(1-p_1)\left[\frac{a^t(z) - a^t(0)}{z}\right] \tag{3.2.3b}$$

$$Eq^{t+1,1}(z) = \sum_{i=1}^{\infty} p_i \left[\frac{q^{t,i}(z)a^t(z) - q^{t,i}(0)a^t(0)}{z}\right] \tag{3.2.3c}$$

$$Eq^{t+1,0} = \left[a^t(0) + p_1(a^t)'(0)\right] q^{t,0} + \sum_{i=1}^{\infty} p_i q^{t,i}(0)a^t(0) \tag{3.2.3d}$$

$$E\tau^{t+1} = (1-p)\tau^t + \frac{1}{\alpha}\left[p_1 q^{t,0}(1-a^t(0)) + \sum_{i=1}^{\infty} p_i q^{t,i}(1)\right] \tag{3.2.3e}$$

where $a^t(z)$ is the generating function for the number of arrivals to each memory node in cycle $[t,t+1)$, and is given by

$$a^t(z) = (1 + \frac{p}{M}(z-1))^{N\tau^t} .$$

Our estimate of the system state at equilibrium is the fixed point $\phi(q,\tau) = (q,\tau)$, which is given by $q = (q^0,q^1,\ldots)$,

$$q^i(z) = r_i\,(a(z))^{i-2}f(z) \quad ; \quad i \geq 2 \tag{3.2.5a}$$

$$q^1(z) = (g(z)-g(0))/z \tag{3.2.5b}$$

$$q^0 = g(0) + 1-\lambda\mu \tag{3.2.5c}$$

$$\tau = \lambda^*/(p\alpha) \tag{3.2.5d}$$

where

$$a(z) = (1+\frac{p}{M}(z-1))^{N\tau} \tag{3.2.6}$$

$$f(z) = (1-\lambda\mu)\frac{a(z)-1}{z-\beta(a(z))}$$

$$g(z) = \frac{f(z)}{a(z)}\beta(a(z)) .$$

where $\beta(z)$ is the probability generating function of the service time distribution.

The only delicate point in the derivation of the expected move operator, (3.2.3), is keeping track of the stage index. For example, (3.2.3e) states that, given the state (q^t,τ^t), at time $t+1$ the average number of

jobs that will be at the think node, $N\tau^{t+1}$, is

- the product of the number at the think node ($N\tau^t$) and the probability of remaining there ($1-p$), plus
- the product of the number of memory nodes holding 0 jobs ($Mq^{t,0}$) and the probability of such a node receiving at least one job and then completing its service ($(1-a^t(0))p_1$), plus
- the sum over $i \geq 1$ of the product of the number of memory nodes with a job at stage i of service ($Mq^{t,i}(1)$) and the probability of such a job completing a service (p_i).

To find the fixed point of the expected move operator, it is more convenient to deal with the state of the system in the *middle* of a cycle: just after arrivals from the think node and before services at the memory nodes. Suppose that $\phi(q,\tau) = (q,\tau)$ is fixed and so is $a(z)$ (given by equation (3.2.6)). Define

$$f^i(z) = q^i(z)a(z) \quad ; \quad i \geq 2$$

$$f^1(z) = q^1(z)a(z) + q^0\left[\frac{a(z)-a(0)}{z}\right]$$

$$f^0 = q^0 a(0)$$

The f^i are the counterparts of the q^i measured at the middle of the cycle instead of the end. The fixed point equation, $\phi(q,\tau) = (q,\tau)$, implies

$$q^i(z) = (1-p_{i-1})f^{i-1}(z) \quad ; \quad i \geq 2 \tag{3.2.7a}$$

$$q^1(z) = \sum_{i=1}^{\infty} p_i \left[\frac{f^i(z)-f^i(0)}{z}\right] \tag{3.2.7b}$$

$$q^0 = \sum_{i=1}^{\infty} p_i f^i(0) + f^0 \ , \tag{3.2.7c}$$

so

$$f^i(z) = (1-p_{i-1})f^{i-1}(z)a(z) \quad ; \quad i \geq 2 \tag{3.2.8a}$$

$$f^1(z) = f^0\left[\frac{a(z)-a(0)}{z}\right] + \sum_{i=1}^{\infty} p_i \left[\frac{f^i(z)a(z)-f^i(0)a(0)}{z}\right] \tag{3.2.8b}$$

$$f^0 = a(0)f^0 + \sum_{i=1}^{\infty} p_i f^i(0)a(0) \ . \tag{3.2.8c}$$

Telescoping (3.2.8a) gives

$$f^i(z) = r_i(a(z))^{i-1} f^1(z) \quad ; \quad i \geq 2 \,, \tag{3.2.9}$$

which combined with (3.2.8b) leads to

$$f^1(z) = \frac{\beta(a(z))f^1(z) - \beta(a(0))f^1(0) + f^0(a(z) - a(0))}{z} \,,$$

since $\beta(z) = \sum_{i \geq 1} r_i p_i z^i = \sum_{i \geq 1} s_i z^i$. Combining (3.2.9) and (3.2.8c) leads to

$$f^0 = \frac{\beta(a(0))f^1(0)}{1 - a(0)} \,,$$

and

$$f^1(z) = \frac{\beta(a(0))f^1(0)}{1 - a(0)} \frac{a(z) - 1}{z - \beta(a(z))} \,,$$

which implies

$$f^1(1) = \frac{\beta(a(0))f^1(0)}{1 - a(0)} \frac{\lambda}{1 - \lambda\mu} \,, \tag{3.2.10}$$

where $\lambda = a'(1)$ and $\mu = \beta'(1)$. By (3.2.9), $f^i(1) = r_i f^1(1)$, so

$$1 = f^0 + \sum_{i=1}^{\infty} f^i(1) = \frac{\beta(a(0))f^1(0)}{1 - a(0)} + \mu f^1(1) \,,$$

which taken together with (3.2.10) gives

$$\frac{\beta(a(0))f^1(0)}{1 - a(0)} = 1 - \lambda\mu \,.$$

Thus, the fixed point equation (3.2.8) becomes

$$f^i(z) = r_i(a(z))^{i-1} f^1(z) \quad ; \quad i \geq 2 \tag{3.2.11a}$$

$$f^1(z) = (1 - \lambda\mu) \frac{a(z) - 1}{z - \beta(a(z))} \tag{3.2.11b}$$

$$f^0 = 1 - \lambda\mu \,. \tag{3.2.11c}$$

(It is not surprising that $f^1(1) = \lambda = a'(1)$ since $f^1(1)$ represents the proportion of memory nodes having a job entering service per cycle.)

To complete the analysis, we need to account for the jobs at the think node. Like (3.2.1), the balance

equation

$$\tau = 1 - \frac{1}{\alpha}\left[\sum_{i=1}^{\infty}(f^i)'(1) + f^i(1) - \lambda\right] .$$

states that the number of jobs at the think node plus the number at the memory nodes equal the total number, N; the quantity λ/α is subtracted from the proportion of jobs at the memory nodes to obtain the proportion at the beginning of the cycle. By (3.2.11b),

$$(f^1)'(1) = \frac{1}{2}\left[\frac{a''(1)}{1-\lambda\mu} + \frac{\lambda^3(\mu_2-\mu)}{1-\lambda\mu}\right] ,$$

where $\mu_2 = \beta''(1)+\beta'(1)$ is the second moment of the service time. Since $\sum_{i\geq 1} f^i(1) = 1-f^0 = \lambda\mu$,

$$\tau = 1 - \frac{1}{\alpha}\left[\frac{\mu}{2}\frac{a''(1)}{1-\lambda\mu} + \frac{\lambda^2(\mu_2-\mu)}{1-\lambda\mu} + \lambda\mu - \lambda\right] .$$

By the definition of $a(z)$, $\lambda = a'(1) = \alpha p \tau$, giving (3.2.5d) and

$$\alpha = \frac{\lambda}{p} + \frac{\mu}{2}\frac{a''(1)}{1-\lambda\mu} + \frac{\lambda^2}{2}\frac{\mu_2-\mu}{1-\lambda\mu} + \lambda(\mu-1) .$$

Equation (3.2.2), which determines $\lambda = \lambda^*$, follows from

$$a''(1) = \lambda^2 - \frac{\lambda p}{M} .$$

Combining (3.2.11) and (3.2.7) leads to (3.2.5).

Table 3 gives the results of experiments that test the accuracy of the approximate analysis, for the case where both think and service time distributions are geometrically distributed, with parameters p and q, respectively. Though the relative error in the estimate never exceeded 3%, the largest errors cropped up when q is small. A small value of q implies, in particular, a large service time variance. In other experiments, we found that the error increases with the service time variance when all other parameters are fixed. This is illustrated in Table 4, where the service time is chosen to be 1 with probability $1 - 1/L$ and $L + 1$, with probability $1/L$, and L is a parameter. Thus, the mean service time is 2, and the variance $L - 1$. As the service time variance becomes large, the throughput drops, because jobs tend to pile up in just a few memory node queues. As the service time variability becomes large with the system size and all other

parameters fixed, the accuracy of the approximation becomes poor.

		Think parameter, p				
		0.2	0.4	0.6	0.8	1.0
Service Parameter, q	0.2	1.346 [−1.868%]	1.559 [−2.295%]	1.634 [−2.437%]	1.671 [−2.543%]	1.694 [−2.717%]
	0.4	2.158 [−0.922%]	2.918 [−1.755%]	3.227 [−2.022%]	3.385 [−2.137%]	3.481 [−2.273%]
	0.6	2.637 [−0.401%]	4.071 [−1.197%]	4.774 [−1.577%]	5.152 [−1.698%]	5.381 [−1.760%]
	0.8	2.933 [−0.134%]	5.025 [−0.670%]	6.265 [−1.058%]	6.980 [−1.176%]	7.417 [−1.147%]
	1.0	3.129 [−0.003%]	5.802 [−0.231%]	7.695 [−0.492%]	8.884 [−0.543%]	9.626 [−0.421%]

Table 3. Equilibrium throughput, measured in simulations, with $N = M = 16$, geometric think time with parameter p, and geometric service time with parameter q. The number in square brackets give the relative error of the estimate (3.2.2). The widths of the 95% confidence intervals computed for the simulation data were never greater than 0.07% of the values in the Table.

Service Time Parameter, L					
2	4	8	16	32	64
4.592 [−1.230%]	4.260 [−1.642%]	3.816 [−2.354%]	3.295 [−3.522%]	2.755 [−5.383%]	2.251 [−8.247%]

Service Time Parameter, L					
128	256	512	1024	2048	4096
1.820 [−12.57%]	1.469 [−18.36%]	1.229 [−27.71%]	1.076 [−39.64%]	0.993 [−52.60%]	0.937 [−64.00%]

Table 4. Equilibrium throughput, measured in simulations, with service times chosen to be 1 with probability $1 - 1/L$ and $L + 1$ with probability $1/L$. The number of processors $N = 16$, the number memory nodes $M = 16$, and the think time parameter $p = 1$ (0 think time). The numbers in square brackets give the relative errors of the analytic estimate, (3.2.2). The widths of the 95% confidence intervals computed for the simulation data were never greater than 0.6% of the values in the Table.

3.3 ACCESS CORRELATIONS MODEL

In this section we consider a model that differs from the basic geometric think time model only in that a job's choice of the next memory node to access may depend on its previous choice. Suppose that the M

memory nodes are partitioned into m classes where the i-th class consists of M_i memory nodes, $\sum_{i=1}^{m} M_i = M$. If a job last accessed a memory node in class i then it next accesses one in class j with probability $\xi_{i,j}$, where $\sum_{j=1}^{m} \xi_{i,j} = 1$. (The first class accessed is arbitrary.) The matrix $(\xi_{i,j})_{1 \le i,j \le M}$ is the transition matrix for the Markov chain describing the sequence of memory classes that any given job visits. For simplicity, we assume this Markov chain is aperiodic and irreducible [4], which implies that for each class there is a steady state probability of a job visiting that class, and this probability is independent of the first class visited.

To represent the system state, group the jobs at the think node by the class selected for next access,

$$N\,\tau^{t,i} = \text{number of jobs at the think node at time } t \text{ whose next access will be to memory class } i\ ,$$

and let

$$M_i\,q_j^{t,i} = \text{number of memory nodes in class } i \text{ that hold } j \text{ jobs at time } t\ .$$

Collecting these quantities in vectors, define

$$\tau^t = (\tau^{t,1}, \ldots, \tau^{t,m})$$

$$q^t(z) = (q^{t,1}(z), \ldots, q^{t,m}(z)) \quad ; \quad \text{where } q^{t,i}(z) = \sum_{j=0}^{\infty} q_j^{t,i}\, z^j$$

The *state* at time t is the pair (q^t, τ^t). Let

$$\mathrm{E}\,\tau^t = (\mathrm{E}\,\tau^{t,1}, \ldots, \mathrm{E}\,\tau^{t,m}) \text{ and } \mathrm{E}q^t = (\mathrm{E}\,q^{t,1}, \ldots, \mathrm{E}\,q^{t,m})\ .$$

Proposition 3.3: *Our estimate of the utilization of a memory node at equilibrium is* λ^*, *the unique root* $\lambda = \lambda^*$ *in* $[0,1]$ *of the polynomial,*

$$\alpha = \frac{\lambda}{p} + \sum_{i=1}^{m} \frac{\lambda c_i(\lambda c_i - p/M)}{2(M_i/M - \lambda c_i)}\ , \tag{3.3.1}$$

where $c = (c_1, c_2, \ldots, c_m)$ *is the invariant measure of the Markov chain* $(\xi_{i,j})$*;*

$$c_i = \sum_{j=1}^{m} c_j\,\xi_{j,i} \quad ; \quad \sum_{i=1}^{m} c_i = 1 \quad ; \quad c_i \ge 0\ ,\ 1 \le i \le m\ . \tag{3.3.2}$$

The quantity $\lambda_i = c_i\lambda^*$ *is an estimate of the equilibrium utilization of a memory node in class i. The*

expected move operator

$$\phi(q,\tau) = \mathrm{E}\left[(q^{t+1},\tau^{t+1}) \mid (q^t,\tau^t) = (q,\tau)\right]$$

is given by,

$$\mathrm{E}\, q^{t+1,i}(z) = \frac{q^{t,i}(z)\, a^{t,i}(z) - q^{t,i}(0)\, a^{t,i}(0)}{z} + q^{t,i}(0)\, a^{t,i}(0) \tag{3.3.3a}$$

$$a^{t,i}(z) = (1 + \frac{p}{M_i}(z-1))^{N\tau^{t,i}} \tag{3.3.3b}$$

$$\mathrm{E}\,\tau^{t+1,i} = (1-p)\,\tau^{t,i} + \frac{1}{N}\sum_{j=1}^{m} M_j(1 - q^{t,j}(0)a^{t,j}(0))\,\xi_{j,i}\ . \tag{3.3.3c}$$

Our estimate of the system state at equilibrium is the fixed point $\phi(q,\tau) = (q,\tau)$, *where* $q = (q^1, \ldots, q^m)$ *and* $\tau = (\tau^1, \ldots, \tau^m)$, *are given by*

$$\tau^i = c_i\lambda^*/(p\alpha) \tag{3.3.4}$$

$$q^i(z) = (1 - \frac{M}{M_i}\, c_i\lambda^*)\left[\frac{z-1}{z - a^i(z)}\right] \quad ; \quad a^i(z) = (1 + \frac{p}{M_i}(z-1))^{N\tau^i}\ , \tag{3.3.5}$$

for $1 \le i \le m$.

The derivation of (3.3.3) is similar to the corresponding derivation for the basic geometric think time model, noting that at cycle $[t,t+1)$ each memory node of class i receives a random number of jobs with the binomial distribution (3.3.3b). Among the class j memory nodes, on average $M_j\,(1-q^{t,j}(0)a^{t,j}(0))$ return a job to the think node in cycle $[t,t+1)$. This gives (3.3.3c), since each such job next accesses class i with probability $\xi_{j,i}$ and each job in the think node at the start of the cycle remains there with probability p.

To find the fixed point $(q,\tau) = \phi(q,\tau)$, suppose that for each i, $1 \le i \le m$,

$$\mathrm{E}q^{t+1,i}(z) = q^{t,i}(z) = q^i(z) \quad ; \quad a^{t,i}(z) = a^i(z) = (1 + \frac{p}{M_i}(z-1))^{N\tau^i}$$

$$\mathrm{E}\tau^{t+1,i} = \tau^{t,i} = \tau^i\ .$$

The analysis of the each component q^i and τ^i is the same as the analysis of their counterparts in the basic geometric think time model. Letting

$$\lambda_i = \frac{M_i}{M}(a^i)'(1) = p\alpha\tau^i ,$$

and evaluating $q^i(1)$, we find that

$$1 - q^i(0)a^i(0) = \frac{M}{M_i}\lambda_i . \tag{3.3.6}$$

Moreover, (3.3.6) and (3.3.3c) imply

$$\lambda_i = \sum_{i=1}^{m} \lambda_j \xi_{j,i} \quad ; \quad 1 \le i \le m .$$

Since the Markov chain with transition matrix $(\xi_{i,j})$ is aperiodic and irreducible, it has a unique invariant measure c satisfying (3.3.2). Hence,

$$\lambda_i = c_i\lambda \quad \text{where } \lambda = \sum_{i=1}^{m} \lambda_i ,$$

which reduces the problem to finding λ. Accounting for the placements of all N jobs gives

$$N = \sum_{i=1}^{m} M_i\,(q^i)'(1) + N\tau^i . \tag{3.3.12}$$

But $\tau^i = \lambda_i/(p\alpha)$ and

$$(q^i)'(1) = \frac{(a^i)''(1)}{2(1 - \lambda_i M/M_i)}$$

$$(a^i)''(1) = \frac{M^2}{M_i^2}\lambda_i(\lambda_i - p/M) ,$$

which taken together with (3.2.12) gives (3.3.1). Taking λ^* as the root of (3.3.1) in $[0,1]$, we obtain $\lambda_i = c_i\lambda^*$, and then (3.3.4) and (3.3.5). We know that (3.3.1) has a unique root in $[0,1]$ because the right hand side is an increasing function of λ. (Consider the Taylor expansion.)

In this model, there are many parameters: the number of processors N, the think time parameter p, the number of classes K, the class sizes M_i, and the class transition probabilities $\xi_{i,j}$. We now report the results of experiments where, for simplicity, we fixed $N=64$, $p=1$, $K=3$, $M_1=4$, $M_2=8$ and $M_3=52$, and defined the class transition matrix in terms of three "stickiness" parameters s_1, s_2 and s_3. After accessing a memory node of class i, a processor's next access is targeted to class i with probability s_i, so the expected

number of returns to class i is $s_i/(1-s_i)$. Again, for simplicity, we insisted that c_i, the long run probability of an access to class j be proportional to $s_i/(1-s_i)$:

$$c_i = \frac{s_i/(1-s_i)}{\sum_{j=1}^{3} s_j/(1-s_j)} ,$$

so the greater the expected number of repeated accesses to a class, the greater the frequency at which the class is accessed. To produce these c_i, we let $\xi_{i,i} = s_i + (1-s_i)s_i/(s_1+s_2+s_3)$ and, for $j \neq i$, $\xi_{i,j} = (1-s_i)s_j/(s_1+s_2+s_3)$.

Table 5 gives the results of four experiments based on these parameter settings, with the s_i set to be

1. proportional to the group sizes,
2. skewed more dramatically than in the first case so that the larger the group the greater its frequency of access,
3. equiprobable, and
4. skewed, reversing the first case, so that the smaller the group the greater its frequency of access.

In all these cases, the error in the estimates derived from (3.3.1) are small.

Class 1		Class 2		Class 3			
size	stickiness	size	stickiness	size	stickiness	throughput	
4	0.05	8	0.10	52	0.65	35.894	[0.151%]
4	0.01	8	0.05	52	0.90	33.962	[-0.100%]
4	0.50	8	0.50	52	0.50	11.662	[0.055%]
4	0.65	8	0.10	52	0.05	4.250	[-0.015%]

Table 5. Equilibrium throughput, measured in simulations, for an $N=64$, $M=64$ size system, with the memory nodes partitioned into three classes. The stickiness parameter is related to the frequency of access, as discussed in the text. The numbers in square brackets give the relative errors of the estimates derived from (3.3.1). The widths of the 95% confidence intervals computed for the simulation data were never greater than 0.04% of the values in the Table.

3.4 COMBINING THE MODELS

Up to now we have described how to adapt the analysis of the memory interference model to handle general think distributions, general memory service time distributions, and Markov correlations between

accesses. In each of these three cases, while we generalized one aspect of the basic geometric think time model, we left the other aspects as they were. In this section we demonstrate that the generalizations can be combined.

In all our models, the equilibrium utilization of a memory node is given as the unique root $\lambda = \lambda^*$ of a polynomial. Consider the model where the service time distribution at a memory node is an arbitrary distribution on the positive integers, where

$$r_i = \Pr(\text{service time} \geq i) \quad ; \quad i \geq 1$$

$$s_i = \Pr(\text{service time} = i) \quad ; \quad i \geq 1$$

$$\mu = \sum_{i=1}^{\infty} i s_i \quad , \quad \mu_2 = \sum_{i=1}^{\infty} i^2 s_i \ .$$

For this model, the equation that determines λ is

$$\alpha = \frac{\lambda}{p} + \frac{\mu}{2}\frac{\lambda(\lambda - p/M)}{1-\lambda\mu} + \frac{\lambda^2}{2}\frac{\mu_2 - \mu}{1-\lambda\mu} + \lambda(\mu - 1) \ ,$$

where $\alpha = N/M$ is the ratio of processors to memories, and p is the parameter of the (geometric) think time distribution. Now, allow the think time distribution to be arbitrary, with

$$R_i = \Pr(\text{think time} \geq i) \quad ; \quad i \geq 0$$

$$S_i = \Pr(\text{think time} = i) \quad ; \quad i \geq 0 \ .$$

$$C = \sum_{i=1}^{\infty} (S_i)^2 / R_i \quad , \quad 1/p = 1 + \sum_{i=1}^{\infty} i \, S_i \ .$$

The new equation for the equilibrium utilization of a memory node becomes

$$\alpha = \frac{\lambda}{p} + \frac{\mu}{2}\frac{\lambda(\lambda - C/M)}{1-\lambda\mu} + \frac{\lambda^2}{2}\frac{\mu_2 - \mu}{1-\lambda\mu} + \lambda(\mu - 1) \ .$$

Suppose that the memories are partitioned into m classes, where class i includes M_i memory nodes ($\sum_{1 \leq i \leq m} M_i = M$), and a job that last accessed a node among class i next accesses one among class j with probability $\xi_{i,j}$, where $(\xi_{i,j})_{1 \leq i,j \leq m}$ is the transition matrix of an aperiodic, irreducible Markov chain. Further, suppose that the memory service time parameters depends on the class of memory node (μ, and μ_2 become μ^i, μ_2^i, $1 \leq i \leq m$), and the think time parameters depend on the class selected for the next

access (p and C become p^i and C^i, $1 \le i \le m$). Let γ be the invariant measure of the controlling Markov chain: $\gamma_i = \sum_{1 \le j \le m} \xi_{i,j}\gamma_j$, $\sum_{1 \le i \le n} \gamma_i = 1$. The equation for the equilibrium utilization of a memory node becomes

$$\alpha = \sum_{i=1}^{m} \left[\frac{\lambda\gamma^i}{p} + \frac{\mu^i}{2} \frac{\lambda\gamma^i(\lambda\gamma^i - C^i/M)}{M_i/M - \lambda\gamma_i\mu^i} + \frac{\lambda^2\gamma_i^2}{2} \frac{\mu_2^i - \mu^i}{M_i/M - \lambda\gamma_i\mu^i} + \lambda\gamma_i(\mu^i - 1) \right].$$

It can be verified that the right hand side is increasing in λ and so has a unique root in $[0,1]$.

3.5 FORK/JOIN MODEL

In the *fork/join* model when a job departs the think node it splits into K mini-jobs, each of which independently accesses a memory node chosen uniformly at random. Here $K \ge 1$ is a parameter. Memory nodes serve mini-jobs just as they serve jobs in the basic geometric think time model: In each cycle each memory node serves exactly one job if one is present. As soon as the last of the K mini-jobs associated with a given job is served, the K mini-jobs recombine and rejoin the think node. We assume that the average think time is $1/p - 1, 0 < p \le 1$.

Consider the following simple example. The number of jobs $N = 1$, the number of mini-jobs per job $K = 2$, and the number of memory nodes $M = 2$. Suppose that at $t = 0$, the job is at the think node, so with probability p it departs the think node in cycle $[0,1)$. Let us assume that it does depart. Further, suppose that both of the mini-jobs access memory node 1. As a result, memory node 1 serves one mini-job in this cycle. Hence, at time $t = 1$, the think node is empty and memory node 1 holds one mini-job. In cycle $[1,2)$ the memory node serves that mini-job, so at time $t = 2$ the job is again at the think node, and the situation is the same as it was at $t = 0$.

We shall present an approximate steady state analysis of this model. In principle, one could carry out the same analytic program used for the other models considered in this paper. However, the system state is now far more complicated. In particular, multivariate generating functions are needed to represent the state. The main result of our analysis is:

Proposition 3.4: *Our estimate of the utilization of a memory node at equilibrium is* λ^*, *the unique root* $\lambda = \lambda^*$ *in* $[0,1]$ *of the following polynomial of degree* $1 + K(K-1)/2$:

$$\alpha = \frac{\lambda}{K}\left[L_K + 1/p - 1\right] \tag{3.4.1}$$

where $\alpha = N/M$ *and*

$$L_0 = 0 \tag{3.4.2a}$$

$$L_i = 1 + \sum_{j=0}^{i} \binom{i}{j} w^j (1-w)^{i-j} L_{i-j} \quad ; \quad i \geq 1 \tag{3.4.2b}$$

$$w = \frac{1-\lambda}{1 - \lambda/2 - p/(2M)} \, . \tag{3.4.3}$$

The quantities w *and* L_K *are estimates of the utilization of a memory node (the equilibrium average proportion of mini-jobs served per memory node per cycle), and average response time (the equilibrium average number of cycles a job requires to have its K mini-jobs served), respectively.*

Here is the idea. First, we obtain (3.4.1) by equating two expressions for the memory node utilization, one of which involves the expected response time per job, L_K. A balance condition gives (3.4.3) for the proportion of mini-jobs at memory that are served per cycle. Last, we make the simplifying assumption that at steady state each mini-job at memory receives service independently with fixed probability w. This leads to the recursion (3.4.2).

To begin, let λ denote the steady state throughput counted in mini-jobs per memory node:

$$\lambda M = \text{expected number of memory nodes utilized in a cycle in steady state} \, .$$

Consider a job that departs the think node in cycle $[t,t+1)$. Suppose that the last of the K associated mini-jobs completes service at memory in cycle $[t+s,t+s+1)$, for some $s \geq 0$. Then the job's response time is defined to be $s+1$. Let L_K denote the expected response time at equilibrium. On average, a job takes $L_K + 1/p - 1$ cycles to complete a circuit form the think node to memory and back. Therefore the throughput counted in jobs per cycle is $N/(L_K + 1/p - 1)$, and the throughput counted in mini-jobs per cycle is

$$\frac{NK}{L_K + 1/p - 1} = \lambda M \, ,$$

which is equivalent to (3.4.1) since $\alpha = N/M$.

Consider a random cycle $[t,t+1)$ at steady state. Let

$$\bar{q}M = \text{expected number of mini-jobs at memory at time } t \ .$$

During this cycle, on average λM mini-jobs arrive to memory and λM depart, making the proportion of mini-jobs served at memory

$$w = \frac{\lambda}{\bar{q} + \lambda} \ . \tag{3.4.4}$$

Approximating the distribution of arrivals to a given memory node as a binomial distribution, we are led to

$$\bar{q} = \frac{\lambda(\lambda - p/M)}{2(1-\lambda)} \ , \tag{3.4.5}$$

which combined with (3.4.4) gives equation (3.4.3) for w. Now, let us introduce the approximation that a mini-job at memory at time t in steady state is served with probability w, independently of every other mini-job at memory. Consider a job at memory at time t at steady state, and assume that exactly i $(1 \le i \le K)$ of its constituent mini-jobs still require service, i.e., i of the mini-jobs await service and $K-i$ have already completed service at memory. By assumption, for each j $(0 \le j \le i)$, j of the i mini-jobs are served in cycle $[t,t+1)$ with probability $\binom{i}{j} w^j (1-w)^{i-j}$. Hence, L_K, the expected number of cycles the whole job takes at memory, satisfies (3.4.2).

We note that having solved (3.4.1)-(3.4.3) for w, we can solve for any moment of a job's response time at memory, via recursions similar to (3.4.3).

Table 6 gives the results of experiments where we varied the number of processors N and the fork value K, keeping other parameters fixed. It is apparent that throughput increases with both N and K. The estimates derived from (3.4.1) are accurate unless N is very small. (The greatest inaccuracies by far occur for $N=1$.) The accuracy generally improves as N increases.

Processors, N	Fork Value, K = 1	2	4	8	16
1	1.000 [0.0%]	1.882 [.080%]	2.966 [4.355%]	3.879 [13.687%]	5.197 [17.060%]
2	1.938 [–0.097%]	3.379 [0.181%]	4.848 [3.301%]	6.494 [3.460%]	8.427 [3.236%]
4	3.627 [–0.265%]	5.653 [0.009%]	7.486 [0.691%]	9.447 [–0.179%]	11.345 [–0.655%]
8	6.315 [–0.427%]	8.546 [–0.344%]	10.371 [–0.628%]	12.027 [–0.899%]	13.379 [–0.887%]
16	9.626 [–0.421%]	11.372 [–0.442%]	12.713 [–0.605%]	13.789 [–0.569%]	14.599 [–0.513%]
32	12.423 [–0.214%]	13.431 [–0.259%]	14.219 [–0.297%]	14.828 [–0.258%]	15.272 [–0.246%]
64	14.147 [–0.069%]	14.654 [–0.102%]	15.071 [–0.111%]	15.395 [–0.105%]	15.630 [–0.126%]

Table 6. Equilibrium throughput, measured in simulations, for an $M = 64$ memory node system, with think time parameter $p = 1$ (0 think time), as a function of the number of processors N and the fork parameter K. The numbers in square brackets give the relative errors of the estimates derived from (3.4.1). The widths of the 95% confidence intervals computed for the simulation data were never greater than .04% of the values in the Table.

4. FINAL REMARKS

A basic, widely used model of memory interference in multiprocessors was considered, along with several generalizations. A simple, unified approach was taken to the approximate analysis of these system, yielding performance estimates that simulations showed to be accurate. The chief weakness of the estimates was in dealing with the case where the variance in the service time is large with respect to the number of processors N and memories M. In the formulae applying to general processor think time distributions, intuitively the term p/M corrects for the finite system size ($1/p - 1$ is the average think time). In the formulae applying to general memory service time distributions, there are no corresponding terms μ/M or μ_2/M (μ and μ_2 are the first two moments of the service time). An approximation that usefully incorporates these "missing" terms would be of great interest. One idea for obtaining better approximations for the general service time case is to take into account the variability in the number of processors thinking. Instead, of treating this number as if were the constant λM, it would be more accurate

to treat it as normally distributed with mean λM and standard deviation $O(\sqrt{M})$.

REFERENCES

[1] Baskett, F. and Smith, A.J., "Interference in multiprocessor computer systems with interleaved memory", *Communications of the ACM*, 19, 6, pp. 327-334 (July 1976).

[2] D. P. Bhandarkar, "Analysis of memory interference in multiprocessors", *IEEE Transactions on Computers*, C-24, 9 (Sept. 1975), 897-908.

[4] Billingsley, P., *Probability and Measure*, second edition, John Wiley & Sons, 1986.

[5] Boguslavsky, L.B., Greenberg, A.G., Jacquet, P., Kruskal, C.P., and Stolyar, A.L., "Models of Memory Interference in Multiprocessors, Part II: Asymptotics and Limit Theorems", submitted.

[6] Fukuda, A., "Equilibrium Point Analysis of Memory Interference in Multiprocessor Systems", *IEEE Transactions on Computers*, C-37, 5 (May 1988), 585-593.

[8] C. H. Hoogendoorn, "A general model of memory interference in multiprocessors", *IEEE Transactions on Computers*, C-26, 10 (Oct. 1977), 998-1005.

[9] "Input versus output queueing on a space-division packet switch", M. J. Karol, M. G. Hluchyj, and S. P. Morgan, *IEEE Transactions on Communications*, COM-35, 12 (Dec. 1987), 1347-1356.

[10] Kleinrock, L., *Queueing Systems*, Volume 1, Wiley, 1975.

[11] B. R. R79, "Interleaved memory bandwidth in a model of a multiprocessor computer system", *IEEE Transactions on Computers*, C-28, 9 (Sept. 1979),

[12] C. V. Rau, "On the bandwidth and interference in interleaved memory systems", *IEEE Transactions on Computers*, C-21, 8 (Aug. 1972), 899-901.

[13] A. S. Sethi and N. Deo, "Interference in multiprocessor systems with localized memory access probabilities", *IEEE Transactions on Computers*, C-28, 2 (Feb. 1979), 157-163.

[14] H. J. Siegel, W. Nation, C. P. Kruskal, and L. M. Napolitano, Jr., "Uses of the Multistage Cube Network Topology", *Proceedings of the IEEE*, vol. 77, no. 12 (Dec. 1989), 1932-1953;

[15] C. Skinner and J. Asher, "Effect on storage contention on system performance", *IBM System Journal*, vol. 8, 319-333, 1969.

[16] B. Smilauer, "General model for memory interference in multiprocessors and mean value analysis", *IEEE Transactions on Computers*, C-34, 8 (Aug. 1985), 744-751.

[18] W. D. Strecker, "An analysis of the instruction execution rate in certain computer structures", Ph.D. dissertation, Carnegie-Mellon Univ., June 1970.

[19] Yen, W.C. and Fu, K.S., "Performance Analysis on Multiprocessor Memory Organizations", ACM Pacific '80, Distributed processing, new directions for a new decade, San Francisco (November 1980), 142-153.

[20] Yen, D.W.L., Patel, J.H., and Davidson, E.S., "Memory interference in synchronous multiprocessor systems." *IEEE Transactions on Computers*, C-31, 11 (Nov. 1982), 1116-1121.

SIMPLE MODELS OF MEMORY INTERFERENCE IN MULTIPROCESSSORS, PART II: ASYMPTOTICS AND LIMIT THEOREMS

Leonid B. Boguslavsky[1], *Albert G. Greenberg*[2], *Philippe Jacquet*[3], *Clyde P. Kruskal*[4], *Alexander L. Stolyar*[1]

December, 1990

Abstract. A basic widely used stochastic model of memory interference in multiprocessors is considered. No useful closed form solutions are known for the key performance measures such as the memory utilization. In our companion paper we provided simple estimates of these performance measures, along with experimental results showing the estimates appear to be accurate. The estimates replace the true system process with a simple, approximating process. Here, we consider the behavior of the model as the number of processors and memory nodes become large, with their ratio tending to a limit. We prove that the system process converges to the approximating process and the corresponding performance estimates converge consistently.

1 Institute of Control Sciences, Profsoyuznaya ul.65, Moscow GSP-312, USSR
2 Room 2C-119, AT&T Bell Laboratories, Murray Hill, NJ, USA 07974
3 INRIA, Rocquencourt, 78153 Le Chesnay Cedex, France
4 Dept. of Computer Science, University of Maryland, College Park, MD 20742 USA

1. INTRODUCTION

A basic, widely used model of memory interference in multiprocessor systems is considered [1,8,11]. There are N processors and M memory nodes. Time is discretized into *cycles* of unit duration, modelling the time a processor uses generating one access to a memory node plus the time a memory node uses serving one access. At the start of a given cycle, a processor is either blocked waiting for its last access to be served , or unblocked. Within the cycle:

1. Each unblocked processor, with fixed probability p, issues an access to a memory node chosen uniformly at random, and with probability $1-p$ issues no access.
2. Next, each memory node serves exactly one access (if it has at least one), thereby unblocking the associated processor.

It is assumed that the random decisions are independent from processor to processor and step to step. A mathematical description of the model is given in the next section.

In our companion paper [4], we presented approximate transient and steady state analyses of this memory interference model, which we referred to as the geometric think time model. Here we present rigorous asymptotic analysis that establishes in a very strong sense that the approximations become exact as the system size (parameters N and M) become large. A well known result of Baskett and Smith [1] is a simple asymptotic formula for the memory utilization. Baskett and Smith use simulations results to show that their "asymptotic result is surprisingly accurate" ([1], pp. 329-330). Later researchers came to believe that their results are "asymptotically exact" (see Rau [7], p. 679; Yen, et al. [13], p. 1118; Karol et al. [6], p. 1351; or Smilauer [9], p. 745). For the first time, we make this statement precise and show that it follows from a stronger limit theorem, Theorem 3.2 below, proving the convergence of the probability measure for the equilibrium state of the memory interference model to a Dirac measure assigning probability 1 to a single state. This limiting state is the unique fixed point of a simple deterministic mapping that approximates the random transition function of the memory interference model. In turn, Theorem 3.2 follows from the *uniform* convergence of the probability measure describing the state of the system at any time $t \geq 1$ to the Dirac measure assigning probability 1 to the state obtained via t iterations of the

deterministic mapping (Theorem 3.1 below).

The ideas behind these limit theorems are rather general (see especially the beginning of Section 3), and we have tried to describe them simply so that the reader might find some use for them in characterizing the behavior of other stochastic systems in the limit as their parameters become large.

The paper is organized as follows. In Section 2 the model is described. In Section 3 the asymptotic framework for analysis of the model is presented, and our main result (Theorem 3.1) is stated. Section 4 begins with a high-level description of the analysis. In subsections 4.1, 4.2, and 4.3 the three main stepping stones towards the proof of Theorems 3.1 and 3.2 are presented, with some proofs relegated to the Appendix. Some of the results established in passing give significant insight into the behavior of the model. In Subsection 4.4, Theorem 3.1 is proven via an application of the Prohorov theorem on weak convergence of probability measures, and Theorem 3.2 follows easily. Finally, in Section 5, we derive several Corollaries of Theorem 3.1, with bearing on the approximate analysis presented in our companion paper [4].

This paper arose from independent work [5] and [10], where similar results were derived using the same basic approach to the asymptotic analysis of the memory interference system. The only prior publication of these results was [10].

2. SYSTEM MODEL

The memory interference model can be viewed as the queueing network depicted in Figure 2.1, consisting of N jobs circulating between M single server queues and a single infinite server or think node. The M single server queues model the M memory nodes. The N jobs model the N processors. Sojourns at the single server queues model memory accesses, and sojourns at the infinite server model local processing. Time is counted in discrete cycles, with the t^{th} cycle assumed to occupy the half closed interval $[t,t+1)$. At each cycle,

1. Independently, each job at the think node is routed with probability p to a memory node chosen uniformly at random, and with probability $1-p$ is left at the think node.

2. Independently, each memory node *now* holding at least one job routes one job back to the think node.

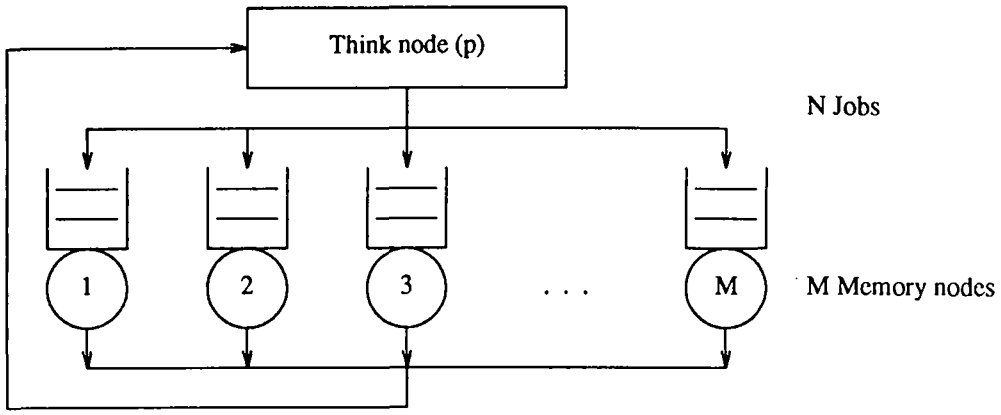

Figure 2.1. Basic geometric think time model.

Our approach to the study of this system is to consider just the number of jobs queued at the memory nodes, not the identities (processor id's) of the jobs queued there. As shown in the companion paper [*BG*+], a wide range of performance statistics can be extracted from this reduced description. Equivalently, we consider just the proportion of the memory nodes that hold i jobs for each $i \geq 0$. It is convenient here and crucial for the theorems to follow to represent the vector of these proportions as a generating function. For $i \geq 0$ and $t \geq 0$, define the generating function

$$q^t(z) = \sum_{i=0}^{\infty} q_i^t z^i$$

where q_i^t denotes the proportion of queues that hold exactly i jobs at time t, i.e., just after the flow of jobs out of the memory nodes in cycle $[t-1, t)$ and just before the flow into the memory nodes in cycle $[t, t+1)$. We call $q^t(z)$ the state generating function or simply the state, and $(q^t)_{t \geq 0}$ the *system process*. The system process is an ergodic Markov chain with a finite state space, which we term $\Omega_{N,M}$. Keep in mind that q^t depends on the parameters N and M. The notation suppresses this dependence because otherwise it would be too unwieldy.

Though this manner of state description is ideal for the analysis, unfortunately there is opportunity for

confusion, so let us recapitulate: The state of the system at time t is represented as a generating function q^t. This generating function qualifies formally as a probability generating function because the coefficients q_i^t are nonnegative and sum to 1: $q^t(1)=1$. However, the coefficients are random variables not probabilities, and q^t is the random state of the system.

As shown in the companion paper, standard generating functions manipulations gives the "expected state" at time $t+1$ as a function of the state at time t: If $q^t(z)=q(z)\in\Omega_{N,M}$, the quantity $Mq'(1)=Mdq/dz(1)$ is the number of jobs queued at memory at time t, so $N-Mq'(1)$ is the number of jobs present at the think node at time t. The number of jobs that depart the think node at the start of cycle $[t,t+1)$ is binomially distributed with parameters $N-Mq'(1)$ and p. The approximate analysis is based on the observation that the formal expectation of the next state q^{t+1} satisfies

$$Eq^{t+1} = \phi(q^t)$$

where we define $Eq^{t+1}(z)=\sum_{i=1}^{\infty} Eq_i^t z^i$, ϕ is the map from generating function to generating function,

$$\phi(q(z)) = \frac{q(z)a(z)-q(0)a(0)}{z}+q(0)a(0) ,$$

and $a(z)=a(q(z))$ is the binomial probability generating function,

$$a(z)=(1+(z-1)p/M)^{N-Mq'(1)} .$$

For large N and M, a simpler approximation is obtained by replacing the binomial generating function with a Poisson counterpart having the same mean. Let $\sigma(z)$ be the map from generating function to generating function given by

$$\sigma(q(z)) = \frac{q(z)e^{\lambda(z-1)}-q(0)e^{-\lambda}}{z}+q(0)e^{-\lambda}$$

where $\lambda = a'(1)$, that is,

$$\lambda=p(N/M-q'(1)) .$$

Note that σ depends on N and M only through their ratio. The reader might recognize σ as the operator describing the transient behavior of an M/D/1 queue (with a state dependent arrival rate λ); that is, if $q(z)$ is the probability generating function describing the probability that the queue length equals i just after a

service completion, for each $i \geq 0$, then $\sigma(q(z))$ is the corresponding probability generating function just after the next service completion, assuming the number of arrivals in between is Poisson distributed with parameter λ.

For small to moderate N and M, the approximations based on ϕ are superior to those based on σ. However, in the asymptotic regime considered here, ϕ becomes indistinguishable from σ, so we deal with σ.

Let $\sigma^t(q(z))$ denote σ applied t times to $q(z)$. Among probability generating functions $q(z)$, σ has a unique fixed point:

$$q^*(z) = \sigma(q^*(z)) \ ;$$

$q^*(z)$ is the generating function of the equilibrium probability distribution for an M/D/1 queue, with fixed arrival rate $\lambda^* = p(N/M - (q^*)'(1))$.

The approximations of the companion paper follow simply from

1. replacing the Markov process $(q^t)_{t \geq 0}$ with the deterministic one $\phi^t(q)$ given $q = q^0 \in \Omega_{N,M}$, and
2. replacing the equilibrium distribution of $(q^t)_{t \geq 0}$ with the fixed point q^* of σ.

We refer to $(\sigma^t(q^0))_{t \geq 0}$ as the *approximating process*. The intuition is that as the parameters N and M become large the system process $(q^t)_{t \geq 0}$ ought to remain near the simpler deterministic process $(\sigma^t(q^0))_{t \geq 0}$, whose transition operator corresponds to the expected move of the system process. Performance estimates can be easily extracted from the approximating process (cf. Section 5), whereas no computational procedure that is even remotely feasible is known for obtaining these estimates from the original Markov process.

3. ASYMPTOTIC FRAMEWORK

We consider the asymptotic regime defined by letting the number of processors N and the number of processors M both become large, with their ratio tending to some fixed constant α. This is the interesting practical limit; it makes sense to increase the memory of a multiprocessor proportionally with the number of processors. Thus, we consider an infinite sequence Γ_α of positive N and an arbitrary function $M = M(N)$

such that

1. $N/M \to \alpha$ as $N \to \infty$, with $N \in \Gamma_\alpha$,

2. $N/M \leq \alpha$ for all $N \in \Gamma_\alpha$.

The second condition is for technical convenience only. It could be dropped at the cost of minor complications in the analysis [10]. Henceforth, we treat the parameter α as fixed, and modify the definition of the operator σ to:

$$\sigma(q(z)) = \frac{q(z)e^{\lambda(z-1)} - q(0)e^{-\lambda}}{z} + q(0)e^{-\lambda} \text{ where } \lambda = p(\alpha - q'(1)) .$$

The notation

$$\lim_{N \in \Gamma_\alpha}$$

will be used as an abbreviation for the limit as $N \to \infty$ with N restricted to Γ_α.

To pose convergence results for the sequence of models $N \in \Gamma_{N,M}$ we need a common state space and a metric for that space. To this end, let R denote a real number > 1, which is determined later as a function of the parameters α and p. Let Ω_α denote the set of probability generating functions (pgf's) with mean $\leq \alpha$ and radius of convergence at least R:

$$\Omega_\alpha = \{q(z) = \sum_{i=0}^{\infty} q_i z^i : q_i \geq 0 , q(1) = 1 , q'(1) \leq \alpha , q(R) < \infty\} .$$

Thus, $\Omega_{N,M} \subset \Omega_\alpha$ for all $N \in \Gamma_\alpha$. For complex valued functions $q(z)$ of a complex variable z, and ρ a positive real number, let

$$\|q\|_\rho = \sup_{|z| \leq \rho} |f(z)| ,$$

For $r, s \in \Omega_\alpha$, define the metric

$$\|r - s\|_\rho$$

for fixed ρ, $1 < \rho < R$. Having established the convergence of pgf's under the $\|\cdot\|_\rho$ metric, we easily obtain fruitful corollaries on the convergence of broad families of functions of the pgf's (Section 5). Though Ω_α is compact with respect to metric $\|\cdot\|_1$, it is not compact with respect to $\|\cdot\|_\rho$ for $\rho > 1$. Lacking

compactness of the entire set, we must work with families of compact subsets; for any $A > 0$, define

$$J_A = \{q(z) = \sum_{i=0}^{\infty} q_i z^i : q_i \geq 0, q(1) = 1, q(R) \leq A\} \tag{3.1}$$

$$K_A = J_A \cap \Omega_\alpha \ . \tag{3.2}$$

Now, let us consider probability measures on the class of Borel sets of Ω_α with the metric $\|\cdot\|_\rho$, for ρ fixed so that $1 < \rho < R$. For each $N \in \Gamma_\alpha$, define

$$\pi_N^t(q) = \Pr(q^t = q) \ ; \ q \in \Omega_{N,M} \ ,$$

and let π_N denote the corresponding equilibrium distribution. To extend π_N^t and π_N to probability measures on Ω_α, define

$$\mu_N^t = \sum_{q \in \Omega_{N,M}} \pi_N^t(q) \delta_q \ ,$$

$$\mu_N = \sum_{q \in \Omega_{N,M}} \pi_N(q) \delta_q$$

where δ_q denotes the Dirac measure concentrated at q:

$$\int \eta d\delta_q = \eta(q) \ \ ; \ \text{for any measurable function } \eta \ .$$

Following Billingsley [2,3], for ξ a probability measure on Ω_α, let $\xi_n \Rightarrow \xi$ denote the weak convergence of a sequence of probability measures $(\xi_n)_{n \geq 0}$ to the probability measure ξ. Now, consider a doubly indexed sequence of probability measures $(\xi_n^m)_{n \geq 0, m \geq 0}$. We say ξ_n^m converges weakly to a probability measure ξ as $n \to \infty$ uniformly in m if, for any bounded continuous function η,

$$\int \eta \ d\xi_n^m \to \int \eta \ d\xi \quad ; \quad \text{as } n \to \infty \text{ uniformly in } m \ .$$

Our main result (Theorem 3.1) is that the probability measure for each state q^t of the system process converges to the Dirac measure concentrated at the state $\sigma^t(q^0)$ of the approximating process, uniformly in $t > 0$. This confirms the intuition alluded to above, and helps to justify the approximations of the companion paper.

Theorem 3.1: Suppose that for some $A > 0$, the initial state $q^0(z) \in K_A$ for every $N \in \Gamma_\alpha$. Then $\mu_N^t - \delta_{\sigma^t(q^0)} \Rightarrow 0$, as $N \in \Gamma_\alpha$ tends to infinity, uniformly for $t \geq 0$.

The strength of Theorem 3.1 is in the uniformity in time $t \geq 0$. Theorem 3.1 means that, for any $\beta > 0$,

$$\lim_{N \in \Gamma_\alpha} \sup_{t \geq 0} \Pr(\| q^t - \sigma^t(q^0) \| > \beta) = 0 .$$

Caution: this should not be confused with a statement about

$$\Pr(\sup_{t \geq 0} \| q^t - \sigma^t(q^0) \| > \beta) ,$$

which cannot tend to zero for all β simply because q^t is not deterministic.

It follows from Theorem 3.1 that the equilibrium measure for the system process converges weakly to the Dirac measure concentrated at the fixed point of the approximating process.

Theorem 3.2: As $N \in \Gamma_\alpha$ tends to infinity,

$$\mu_N - \delta_{q^*} \Rightarrow 0 .$$

Note that while some condition on the start state q^0 is unavoidable for showing convergence at specific times t, no such condition is needed in Theorem 3.2, which establishes convergence at equilibrium. An important performance statistic is the equilibrium utilization of a given memory node. In the companion paper, we presented the approximation for the utilization: the root λ in $[0,1]$ of the quadratic equation

$$N/M = \frac{\lambda}{p} + \frac{\lambda (\lambda - p/M)}{2(1 - \lambda)} .$$

In Section 5, we show that this and a large family of related approximations are asymptotically exact.

4. ANALYSIS

To prove Theorem 3.1, we first need a pair of results that we refer to as "expected move theorems" (Theorems 4.1.1 and 4.1.3). Roughly, these two results show that, with probability tending to one as N tends to infinity, q^{t+1} can be found in an arbitrarily small neighborhood of Eq^t. Moreover, this probability tends to zero *uniformly* for $q = q^t$ in certain compact subsets of Ω_α. Besides being stepping stones towards the main goal, the expected move theorems show that as N becomes large the sequence of states of $(q^t)_{t \geq 0}$ assumed over any finite time interval becomes indistinguishable from the corresponding sequence of states of the approximating process $\sigma^t(q^0)_{t \geq 0}$.

Second, we need to show that the deterministic operator σ acts a contraction map (Theorem 4.3.1); for q in any of a large family of compact subsets of Ω_α, $\sigma^t(q) \to q^*$ as $t \to \infty$. This is natural. Though the expected move theorems show that the system process tracks the approximating process over finite time, we wouldn't expect the system process to converge to the fixed point $q^* = \sigma(q^*)$ if the approximating process didn't. The proof of the appropriate contraction property is in the Appendix; it may be surprising that this proof relies on stochastic arguments.

To prove the main result, the key tool is the Prohorov theorem [2] on weak convergence of measures. Using the Prohorov theorem, we can combine the expected move theorem and the contraction property of σ to obtain Theorem 3.1. The main difficulty is in showing that the premise for using the Prohorov theorem holds, namely that the sequence of measures (μ_N^t) is tight. Coming back to the original problem, we establish tightness using a single Markov process that *for every* $N \in \Gamma_\alpha$ stochastically dominates the Markov process $(q^t)_{t \geq 0}$. This is the basis of Theorem 4.2.4. The central technical point (Lemma 4.2.1) is that, after a fixed number of cycles, the rate at which new accesses arrive to a given memory node becomes bounded below 1, for all sufficiently large $N \in \Gamma_\alpha$, no matter what the initial state $q \in \Omega_{N,M}$.

To summarize, the stepping stones to the main result are: (i) the expected move theorems showing uniform convergence over finite time, (ii) the uniform contraction of the deterministic map, (iii) the tightness of the sequence of measures (μ_N^t), and (iv) the application of Prohorov's theorem. We believe this proof strategy may find many other applications.

4.1 EXPECTED MOVE THEOREMS

This section is devoted to proving that given any state $q^t = q \in \Omega_{N,M}$ the next state q^{t+1} becomes arbitrarily close to the "expected move" $Eq^{t+1} = \sigma(q)$ with arbitrarily high probability as $N \to \infty$, uniformly for all states q in $\Omega_{N,M}$, $N \in \Gamma_\alpha$.

For $r, s \in \Omega_\alpha$, define the metric

$$\text{dist}(r, s) = \|r - s\|_1 + |r'(1) - s'(1)| .$$

Theorem 4.1.1 (first expected move theorem): For any $\beta > 0$, there is a sequence $(\varepsilon_N)_{N \in \Gamma_\alpha}$ such that $\lim_{N \in \Gamma_\alpha} \varepsilon_N = 0$ and, for all $N \in \Gamma_\alpha$,

$$\sup_{q \in \Omega_{N,M}} \Pr\{\mathrm{dist}(q^{t+1}, \sigma(q^t)) \geq \beta \mid q^t = q\} < \varepsilon_N, \quad \textit{for all } N \in \Gamma_\alpha . \tag{4.1.1}$$

Remark: The uniformity alluded to above is crucial; it is captured by taking the supremum over all $q \in \Omega_{N,M}$. We also need a corresponding result (Theorem 4.1.3 given immediately below) for the $\|\cdot\|_\rho$ metric. Though the $\|\cdot\|_\rho$ metric is stronger, the second result is not stronger than the first because it establishes uniform convergence over the smaller sets $\Omega_{N,M} \cap K_A$.

Theorem 4.1.1 implies that the stochastic process $(q_t)_{t \geq 0}$ converges to the approximating process $(\sigma^t(q^0))_{t \geq 0}$ over any finite time interval.

Corollary 4.1.2: For all $T \geq 1$ and all $\beta > 0$,

$$\lim_{N \in \Gamma_\infty} \sup_{q \in \Omega_{N,M}} \Pr(\exists t \in [0, T) : \mathrm{dist}(q^t, \sigma^t(q^0)) > \beta \mid q^0 = q) = 0 .$$

Proof of Corollary 4.1.2: It suffices to show that the mapping σ is uniformly continuous on the set Ω_α, with respect to the $\mathrm{dist}(\cdot,\cdot)$ metric. To this end, let us define the larger set

$$L_\alpha = \{(q,x) : q(z) = \sum_{i=0}^{\infty} q_i z^i, q_i \geq 0, q(1) = 1, q'(1) \leq x \leq \alpha\} .$$

Identify the set Ω_α with the subset of L_α of elements (q,x) such that $q'(1) = x$. Let us define the metric $\mathrm{dist}(\cdot,\cdot)$ on L_α

$$\mathrm{dist}((q,x),(h,y)) = \|q - h\|_1 + |x - y| ,$$

and note this definition coincides with the earlier one on Ω_α. The quantity $q'(1)$, regarded as a real valued function of q, is lower semi-continuous, with respect to the metric $\|\cdot\|_1$. This fact and the Arzella-Ascolli Theorem [2] imply that the set L_α is compact with respect to the $\mathrm{dist}(\cdot,\cdot)$ metric

Next, define the mapping σ on L_α by

$$\sigma(q,x) = (S_\lambda q, (1-p)x + p(1 - q(0)e^{-\lambda})) ,$$

where $\lambda = p(\alpha - x)$ and

$$S_\lambda q(z) = \frac{q(z)e^{\lambda(z-1)} - q(0)e^{-\lambda}}{z} + q(0)e^{-\lambda} .$$

This definition of σ coincides with the earlier one on the set Ω_α. Clearly, the mapping σ is continuous with respect to the dist$(\cdot,\cdot)$ metric. Last, since L_α is compact, the mapping σ is uniformly continuous. ■

Proof of Theorem 4.1.1: First let us drop superscripts: $q^t = q$, $q^{t+1} = s$. Break the transitions $q \to s$ and $q \to \sigma(q)$ into two steps:

$$q \to r \to s$$

$$q \to q(z)e^{\lambda(z-1)} \to \sigma(q) \; ; \quad \lambda = p(\alpha - q'(1))$$

where the intermediate state r represents the state of the system within cycle $[t,t+1)$ just after jobs have departed the think node and have joined queues at the memory nodes, and just before the memory nodes serve jobs for this cycle. Specifically, letting Q_i denote the queue length at memory node i at time t, and A_i the number of arrivals to that node in cycle $[t,t+1)$,

$$q(z) = \frac{1}{M} \sum_{i=1}^{M} z^{Q_i}$$

and

$$r(z) = \frac{1}{M} \sum_{i=1}^{M} z^{Q_i + A_i} .$$

It is not hard to show that if pgf's r and $q(z)e^{\lambda(z-1)}$ are close (with respect to the metric dist$(\cdot,\cdot)$) then so are pgf's s and $\sigma(q)$. That is, it can be verified that (4.1.1) holds if there is a sequence $(\varepsilon_N)_{N \in \Gamma_*} \to 0$ such that

$$\sup_{q \in \Omega_{N,M}} \Pr(\| r - q(z)e^{\lambda(z-1)} \|_1 > \beta) \le \varepsilon_N \tag{4.1.2}$$

$$\sup_{q \in \Omega_{N,M}} \Pr(|r'(1) - q'(1) - \lambda| > \beta) \le \varepsilon_N . \tag{4.1.3}$$

(Note $r(z)$ is a random function of $q(z)$.)

Let τM be the number of jobs at think node at the beginning of cycle $[t,t+1)$, so $\tau = N/M - q'(1)$. Define the generating function $m(z)$ by

$$m(z) = \left[1 - \frac{p}{M} + \frac{pz}{M}\right]^{M\tau} .$$

Elementary exponential approximation gives

$$m(z) = e^{(z-1)p\tau/M}\,(1 + O(\frac{1}{M}))\,.$$

Since $N/M \to \alpha$, it can be shown that (4.1.2) holds if there is a sequence $(\varepsilon_N)_{N \in \Gamma_*} \to 0$ such that

$$\Pr(\|r(z)-q(z)m(z)\|_1 > \beta) \le \varepsilon_N\,, \tag{4.1.4}$$

uniformly for $q \in \Omega_{N,M}$. Now let us prove (4.1.4). The (simpler) proof of (4.1.3) will be sketched afterwards.

All bounds to follow are uniform in q and $\tau \le \alpha$. Let $\bar{z}$ denote the complex conjugate of z. For fixed z, define the complex random variable

$$A_i(z) = z^{A_i}\,,$$

and let

$$v(z) = \left|1 - \frac{p}{M} + \frac{p|z|^2}{M}\right|^{M\tau} - \left|1 - \frac{p}{M} + \frac{pz}{M}\right|^{2M\tau}\,,$$

$$c(z) = \left|1 - \frac{2p}{M} + \frac{p(z+\bar{z})}{M}\right|^{M\tau} - \left|1 - \frac{p}{M} + \frac{pz}{M}\right|^{2M\tau}\,.$$

The complex functions $m(z)$, $v(z)$, and $c(z)$ may be regarded as formal definitions of the expected value of $A_i(z)$, the variance of $A_i(z)$, and the covariance of $A_i(z)$ and $A_j(z)$, for any $1 \le i,j \le M, i \ne j$. Since $q(z)$ is fixed,

$$\mathrm{E}r(z) = \frac{1}{M}\sum_{i=1}^{M} z^{Q_i}\mathrm{E}[z^{A_i}] = q(z)m(z)\,,$$

and (4.1.4) becomes

$$\Pr(\|r(z) - \mathrm{E}r(z)\|_1 > \beta) \le \varepsilon_N\,. \tag{4.1.5}$$

By elementary exponential approximation,

$$v(z) = e^{(|z|^2-1)p\tau}(1 + O(\frac{p^2\tau}{M})) - |e^{(z-1)p\tau}|^2\,(1 + O(\frac{p^2\tau}{M}))\,,$$

$$c(z) = e^{(z+\bar{z}-2)p\tau}\,O(\frac{p^2\tau}{M})\,,$$

uniformly for z in any compact subset of the complex plane, in particular, for z in the unit disk. Again,

since $q(z)=\sum_i z^{Q_i}$ is fixed, the variance of the random state $r(z)$ is (formally) given by

$$\begin{aligned}\mathrm{var}(r(z)) = \mathrm{var}\Big(\frac{1}{M}\sum_i z^{Q_i+A_i}\Big) &= \frac{1}{M^2}\left[\sum_i |z|^{2Q_i}\, v(z) + \sum_{i\neq j} z^{Q_i}\bar{z}^{Q_j}\, c(z)\right] \\ &= \frac{1}{M}\, q(|z|^2)\, v(z) + \left[|q(z)|^2 - \frac{q(|z|^2)}{M}\right] c(z) \\ &= O(\frac{1}{M})\,,\end{aligned}$$

uniformly for $|z| \le 1$. By Chebychev's inequality, for any $\theta > 0$ and $|z| \le 1$,

$$\Pr(|r(z)-Er(z)| > \theta) \le \frac{\mathrm{var}(r(z))}{\theta^2} \le O(\frac{1}{M})\frac{1}{\theta^2}\,. \tag{4.1.6}$$

This last inequality would establish (4.1.5) if $|\cdot|$ were replaced with $\|\cdot\|_1$; take $\theta=\beta M^{-1/4}$. To justify such a replacement we shall weaken the bound a little. Let L be one plus the integer part of $\sqrt{M}$, and choose $z_1, \ldots, z_L$ on the unit circle such that, for every z on that circle

$$\min_{k\le L} |z-z_k| \le 10\sqrt{M}\,.$$

For example, take $z_k=\exp(2ik\pi/\sqrt{M})$. Then

$$\begin{aligned}\Pr\Big[\sup_{1\le k\le L} |r(z_k)-Er(z_k)| \le \theta\Big] &\ge 1 - \sum_{k=1}^{L} \Pr(|r(z_k) - Er(z_k)| > \theta) \\ &= 1 - L\,O\left[\frac{1}{M}\right]\frac{1}{\theta^2} \\ &\ge 1 - O(M^{-1/2})\frac{1}{\theta^2}\,.\end{aligned} \tag{4.1.7}$$

On the unit disk, the derivatives of $r(z)$ and $Er(z)$ are maximal at $z=1$, and are less than or equal to N/M there. Hence, for every z on the unit circle there is a k ($1 \le k \le L$), such that

$$|r(z) - Er(z) - r(z_k) + Er(z_k)| \le 2\,\frac{N}{M}\,|z-z_k| \le 20\,\frac{N}{M}\,M^{-1/2}\,. \tag{4.1.8}$$

Combining (4.1.7) and (4.1.8) then shows that, for any $\theta > 0$,

Proof: If q^0 is in a compact set K_A then with probability tending to one every q^t for $t \in [0,T]$ belongs to the compact set K_B for some $B > A$. The rest of the proof is the same as the proof of Corollary 4.1.2, exploiting the uniform continuity of σ on K_B. ■

4.2 TIGHTNESS

To argue about the weak convergence of the family of measures (μ_N^t) we will need to establish that the family is tight; i.e., as either $N \in \Gamma_\alpha$ or $t \geq 0$ tends to infinity, the mass of μ_N^t does not "escape to infinity", but instead remains concentrated in compact subsets of Ω_α.

Let us define the *intensity* of the arrivals to a given memory node at the t^{th} cycle as the random variable

$$\lambda^{t+1} = p(N/M - (q^t)'(1)) \ .$$

Thus, the number of jobs at the think node at time t is $M\lambda^{t+1}/p$. If $N/M > 1$ then the initial intensity λ^0 may be greater than 1. However, after a fixed number of steps a the intensity λ^a must dip below 1, by the following analysis of the approximating process and (expected move) Theorem 4.1.1.

Lemma 4.2.1: For some positive integer, a, and $0 < \delta < 1$,

$$\lim_{N \in \Gamma_\alpha} \sup_{q \in \Omega_{N,M}} \Pr(\lambda^a \geq 1 - \delta \mid q^0 = q) = 0 \ .$$

Proof: The counterpart of λ^{t+1} for the approximating process $(\sigma^t(q^0))_{t \geq 0}$ is $\nu^{t+1} = p(\alpha - (\sigma^t(q^0))'(1))$. By Corollary 2.1, for any fixed $a > 0$, as $N \in \Gamma_\alpha$ tends to infinity, $|\lambda^a - \nu^a|$ tends to zero with probability tending to one, uniformly for $q \in \Omega_{N,M}$. Thus, we need only show that there exist a and $0 < \delta < 1$ such that

$$\nu^a \leq 1 - \delta \ ; \quad \text{uniformly for } q \in \Omega_\alpha \ . \tag{4.2.1}$$

Let $q^0(z) = q(z) = \sum_{i \geq 0} q_i z^i \in \Omega_\alpha$. By the observation that $(\sigma^T q)(0)$ depends only on the terms $q_0, q_1, \ldots, q_{T+1}$, it is easily shown that with $T = 2\lfloor \alpha \rfloor$

$$\begin{aligned} (\sigma^T q)(0) &\geq e^{-T\alpha}(q_0 + q_1 + \cdots + q_T) \\ &\geq e^{-T\alpha}/2 \ , \end{aligned}$$

since $q'(1) \leq \alpha$. Taking $\varepsilon = e^{-T\alpha}/2$, for all $t > T$,

$$(\sigma^t q)(0) = (\sigma^T \sigma^{t-T} q)(0) \geq \varepsilon$$

since $\sigma^{T-t}(q) \in \Omega_\alpha$. For $i \geq 1$ and $t = T+i$,

$$\begin{aligned} \nu^{t+1} &= (1-p)\nu^t + p(1-q^t(0)\ e^{-\nu^t}) \\ &\leq (1-p)\nu^t + p(1-\varepsilon\ e^{-\alpha}) \\ &\leq (1-p)^{i+1}\nu^{t-i} + p(1-\varepsilon\ e^{-\alpha})[(1-p)^i + \cdots + (1-p)^0] \\ &\leq (1-p)^{i+1}\alpha + 1-\varepsilon\ e^{-\alpha}\ . \end{aligned}$$

which establishes (4.2.1). ■

Corollary 4.2.2: There exists an integer b such that

$$\lim_{N \in \Gamma_\alpha} \sup_{q \in \Omega_{N,M}} \Pr\left(\frac{\lambda^0 + \cdots \lambda^{a+b}}{a+b} \geq 1-\delta \mid q^0 = q\right) = 0\ .$$

We are now in position to construct a certain G/D/1 queue that stochastically dominates the queue length process at any given memory node, say, $(Q_1(t))_{t \geq 0}$. This process is used below to establish the tightness of the measures μ_N^t.

Lemma 4.2.3: For some numbers $R > 1, \theta^*$ and B^*,

$$E[q^t(R)] < \theta^* q^0(R) + B^* \quad ; \quad t \geq 0\ .$$

Proof: As before, let $Q_i(t)$ denote the number of jobs at memory node i at time t, so

$$E[q^t(R)] = E[1/M \sum_{i=1}^{M} R^{Q_i}] = E[R^{Q_1(t)}]\ .$$

We show that there is an ergodic Markov process $\bar{Q}(t)$ that can be constructed on the same probability space as $Q_1(t)$ such that, assuming $\bar{Q}(0) = Q_1(0)$, (i) for all sufficiently large $N \in \Gamma_\alpha$, $\bar{Q}(t)$ stochastically dominates $Q_1(t)$ for all $t > 0$, and (ii) for some numbers $R > 1, \theta^*$ and B^*,

$$E[R^{\bar{Q}(t)}] < \theta^* R^{\bar{Q}(0)} + B^* \quad ; \quad t \geq 0\ .$$

First, define $(Q^*(t))_{t \geq 0}$ as the queue length process for the following variation of the G/D/1 queue, with integer parameters $a, b \geq 0$, and real parameter parameters $\varepsilon > 0$, $0 < \beta < 1$, $0 < \delta < 1$. Divide time into consecutive intervals of length $a+b$. Assume that (i) at the first step of each interval a random number of

jobs, A^*, arrives to the queue, and (i) during the remaining $a+b-1$ steps no jobs arrive to the queue. The random variable A^* is independent from interval to interval, and is distributed as follows: with probability $1-\beta$, A^* is Poisson with rate $(a+b)(1-\delta)$, and with probability β, A^* is Poisson with rate $(a+b)p\alpha$. By application of Corollary 4.2.2, it is easy to tune β so that (i) $EA^* < (a+b)$ and (ii) N is large enough that

$$\sup_{q \in \Omega_{N,M}} \Pr\{\lambda^0 + \cdots + \lambda^{a+b} \geq (a+b)(1-\delta)\} \leq \beta \ .$$

Therefore, for each t, the quantity $\lambda^t + \cdots + \lambda^{t+a+b}$ is stochastically dominated by A^*. The role of the parameter $\varepsilon > 0$ is to compensate for the fact that $Q_1(t)$ sees binomially distributed arrivals, which in $Q^*(t)$ are dominated by Poisson ones with slightly greater intensities.

We now use a simple fact about queues with deterministic, unit service times. Consider such a queue and any interval of time $[u,v)$ during which the queue is never empty, and let w denote the number of jobs that arrive during the interval. Changing all w arrival times to u, the start of the interval, can only increase the queue length throughout the interval. If w is replaced by $w' \geq_{st} w$, then throughout the interval the queue length becomes stochastically greater. Returning to the system at hand, it follows that $Q_1(t)$ is stochastically dominated by $a+b+Q^*(t) = \bar{Q}(t)$.

Now let $\bar{\alpha}$ and $\bar{\delta}$ respectively denote $p\alpha+\varepsilon$ and $\delta-\varepsilon$. Let $f^t(z)$ denote the pgf of $Q^*(t)$, formally

$$f^t(z) = E[z^{Q^*(t)}] \ .$$

The transition from $f^{(a+b)t}$ to $f^{(a+b)(t+1)}$ can be described via the application of linear operators defined on pgf's. For any $\lambda > 0$ and any pgf $f(z)$ let

$$(S_\lambda(f))(z) = \frac{f(z)e^{\lambda(z-1)} - f(0)e^{-\lambda}}{z} + f(0)e^{-\lambda} \ .$$

Then

$$f^{(a+b)(t+1)} = (S_0)^{a+b-1}(\beta S_{(a+b)\bar{\alpha}} + (1-\beta)S_{(a+b)(1-\bar{\delta})})(f^{(a+b)t}) \ .$$

Elementary bounds establish the existence of $B > 0$ such that, for any $R \geq 1$,

$$f^{(a+b)(t+1)}(R) \geq \theta(R) f^{(a+b)t}(R) + B \ ,$$

with

$$\theta(r)=\beta\exp((a+b)p\bar{\alpha}(R-1))/R^{a+b}+(1-\beta)\exp((a+b)(1-\bar{\delta})(R-1))/R^{a+b} .$$

It is easy to find $R>1$ such that $\theta(R)<1$. (Note that $\theta(1)=1$ and the first derivative $\theta'(1)<0$.) Therefore, for any integer $t\geq 0$,

$$f^{(a+b)t}(R)\leq(\theta(R))^t f^0(R)+B/(1-\theta(R)) .$$

This establishes the result for those t that are multiples of $a+b$. The proof is easily completed by applying intermediate operators to f^t for any other t. ■

Theorem 4.2.4: If, for some $R>1$, $q^0(R)$ is uniformly bounded for $N\in\Gamma_\alpha$ then the family of measures, (μ_N^t) for $N\in\Gamma_\alpha$ and $t\geq 0$, is tight; that is

$$\lim_{A\to\infty}\int_{K_A} d\mu_N^t = 1 \quad ; \quad \text{uniformly for } N\in\Gamma_\alpha\ ,\ t\geq 0\ ,$$

where the compact sets $K_A, A>1$, are defined in (3.2).

Proof: By Markov's inequality, it suffices to prove that there exists a constant A^* such that, for all N and t,

$$\int q(R)\, d\mu_N^t \leq A^* .$$

By the previous Lemma,

$$\int q(R)\, d\mu_N^t = Eq^t(R) \leq \theta^*\, q^0(R) + B^* .$$

But $q^0(R)$ is uniformly bounded by assumption. ■

4.3 CONTRACTION

In this very short subsection, we state a result showing successive iterations of σ maps any of the compact sets K_A into successively smaller balls centered at the fixed point $q^*=\sigma(q^*)$. The proof is given in the Appendix.

Theorem 4.3.1: For some $R>1$, any $1<\rho<R$, and any $A>0$,

$$\lim_{t\to\infty}\|\sigma^t(q)-q^*\|_\rho=0 \quad ; \quad \text{uniformly for } q\in K_A\ ,$$

where q^* is the unique fixed point of σ.

4.4 PROOFS OF THEOREMS 3.1 AND 3.2

We are now ready to combine the expected move Theorems (Theorems 4.1.1 and 4.1.3), the tightness property (Theorem 4.2.4), and contraction property (Theorem 4.3.1) to prove the main result (Theorem 3.1). Each of the results established thus far for some $R > 1$ continues to hold if R is decreased, provided R remains greater than 1. Henceforth fix $R > 1$ sufficiently small that the results hold simultaneously.

Fix $A > 0$ and suppose that for each $N \in \Gamma_\alpha$, $q^0 = s_N \in K_A$. By Corollary 4.1.4, as $N \to \infty$, $N \in \Gamma_\alpha$, for t fixed,

$$\mu_N^t - \delta_{\sigma^t(s_N)} \Rightarrow 0 . \tag{4.4.1}$$

To prove Theorem 3.1 it suffices to show that the convergence is uniform in $t \geq 0$. We do this via an application of the Prohorov theorem.

Following Billingsley [2,3], for ξ a probability measure, let $\xi\sigma^{-1}$ denote the probability measure determined by

$$\int \eta d\xi\sigma^{-1} = \int \eta \circ \sigma d\xi \quad ; \quad \text{for any measurable function } \eta .$$

Lemma 4.4.1: If $q^0 \in K_A$ for some $A > 0$ then, for every finite k, $\mu_N^{t+k} - \mu_N^t \sigma^{-k} \Rightarrow 0$ as $N \in \Gamma_\alpha$ tends to infinity, uniformly for $t \geq 0$.

Proof: We shall prove this result for $k = 1$. For greater k, the result follows from the $k = 1$ proof and the continuity of σ. Let η be an arbitrary bounded, uniformly continuous function defined on Ω_α. It suffices to show

$$\lim_{N \in \Gamma_\alpha} \left(\int \eta \, d\mu_N^{t+1} - \int \eta \circ \sigma d\mu_N^t \right) = 0 \quad ; \quad \text{uniformly for } t \geq 0 .$$

Theorem 4.1.3 (second expected move theorem) implies that, for each $A > 0$,

$$\lim_{N \in \Gamma_\alpha} \sum_{q' \in \Omega_{N,M}} \Pr(q^{t+1} = q' \mid q^t = q)\eta(q') = \eta(\sigma(q)) ,$$

uniformly for all $q \in \Omega_{N,M} \cap K_A$ and t.

Since the family of probability measures (μ_N^t) is tight with respect to the compact sets K_A, $A \geq 0$, the

following limits hold uniformly for $N \in \Gamma_\alpha$ and $t \geq 0$:

$$\lim_{A \to \infty} \sum_{q \in K_A \cap \Omega_{N,M}} \pi_N^t(q)\, \eta(\sigma(q)) = \int \eta \circ \sigma \, d\mu_N^t$$

$$\lim_{A \to \infty} \sum_{q \in K_A \cap \Omega_{N,M}} \pi_N^t(q) \sum_{q' \in \Omega_{N,M}} \Pr(q^{t+1} = q' \mid q^t = q)\eta(q') = \int \eta \, d\mu_N^{t+1} .$$

Combining the last three limits completes the proof. ■

Lemma 4.4.2: If $q^0 \in K_A$ for some $A > 0$ then the sequence of probability measures $(\mu_N^t \sigma^{-k})_{k \geq 0}$ converges weakly as $k \to \infty$ to the Dirac measure δ_{q^*}, uniformly for $N \in \Gamma_\alpha$ and $t \geq 0$.

Proof: Let η be an arbitrary bounded continuous function on Ω_α. It suffices to show

$$\lim_{k \to \infty} \int \eta \circ \sigma^k d\mu_N^t = \eta(q^*) \quad ; \quad \text{uniformly for } N \in \Gamma_\alpha \, , \, t \geq 0 .$$

By Theorem 4.2.4, (μ_N^t) is a tight family of measures satisfying

$$\lim_{A \to \infty} \int_{K_A} \eta \circ \sigma^k d\mu_N^t = \int \eta \circ \sigma^k d\mu_N^t \quad ; \quad \text{uniformly for } N \in \Gamma_\alpha \, , \, t \geq 0 .$$

On the other hand, for each $A > 0$, as $k \to \infty$, $\sigma^k(q) \to q^*$, uniformly for $q \in K_A$. This fact and the fact that η is uniformly continuous on the compact set K_A gives

$$\lim_{k \to \infty} \int_{K_A} \eta \circ \sigma^k d\mu_N^t = \eta(q^*) \int_{K_A} d\mu_N^t \quad ; \quad \text{uniformly for } N \in \Gamma_\alpha \, , \, t \geq 0 .$$

But the integral on the righthand side converges to 1 as $A \to \infty$. ■

Now, suppose (4.4.1) does not hold uniformly in t. Then there exists $\varepsilon > 0$, a bounded continuous function η, and an infinite sequence $(N(i), t(i))_{i \geq 0}$ for $N(i) \in \Gamma_\alpha, N(i) \to \infty, t(i) \geq 0, t(i) \to \infty$, such that

$$\left| \int \eta \, d\mu_{N(i)}^{t(i)} - \eta(\sigma^{t(i)}(s_{N(i)})) \right| > \varepsilon. \tag{4.4.2}$$

Since μ_N^t is a tight family of measures it follows from Prohorov's theorem that $(N(i), t(i))_{i \geq 0}$ contains a weakly convergent subsequence. For notational convenience, relabel indices to identify $(N(i), t(i))_{i \geq 0}$ with such a subsequence and let ξ denote the corresponding weak limit. Since $s_{N(i)} \in K_A$ and $t(i) \to \infty$ as $i \to \infty$,

$$\eta(\sigma^{t(i)}(s_{N(i)})) \to \eta(q^*) \quad \text{as } i \to \infty$$

by the uniform contraction property established in Theorem 4.3.1. Thus, by (4.4.2),

$$\left| \int \eta \, d\xi - \eta(q^*) \right| > \varepsilon. \tag{4.4.3}$$

By Lemma 4.4.1, for every k, as $i \to \infty$,

$$\mu_{N(i)}^{t(i)} - \mu_{N(i)}^{t(i)-k} \sigma^{-k} \Rightarrow 0 .$$

Hence, for every $k \geq 0$,

$$\mu_{N(i)}^{t(i)-k} \sigma^{-k} \Rightarrow \xi \quad \text{as } i \to \infty . \tag{4.4.4}$$

By Lemma 4.4.2,

$$\mu_N^t \sigma^{-k} \Rightarrow \delta_{q^*} \quad \text{as } k \to \infty \text{ uniformly for } N \in \Gamma_\alpha \, , \, t \geq 0 . \tag{4.4.5}$$

Fix k sufficiently large that, by (4.4.5),

$$\left| \int \eta \, d\mu_{N(i)}^{t(i)-k} \sigma^{-k} - \eta(q^*) \right| < \frac{\varepsilon}{2} \text{ uniformly for } i \geq 0 . \tag{4.4.6}$$

Next, fix i sufficiently large that $t(i) \geq k$ and, by (4.4.4),

$$\left| \int \eta \, d\mu_{N(i)}^{t(i)-k} \sigma^{-k} - \int \eta \, d\xi \right| < \frac{\varepsilon}{2} \tag{4.4.7}$$

Finally, adding (4.4.6) and (4.4.7) contradicts (4.4.3), proving that (4.4.1) does hold uniformly in t, thereby completing the proof of Theorem 3.1.

To establish Theorem 3.2, suffices to show that for any bounded, continuous function η,

$$\lim_{N \in \Gamma_\alpha} \int \eta \, d\mu_N = \eta(q^*) .$$

Since the basic process is an ergodic Markov chain, we may assume $q^0 = 1 \in K_A$, for all $A > 1$, without affecting the invariant measure μ_N, which satisfies

$$\lim_{t \to \infty} \int \eta \, d\mu_N^t = \int \eta \, d\mu_N .$$

By Theorem 3.1,

$$\lim_{t \to \infty} \lim_{N \in \Gamma_*} \int \eta \, d\mu_N^t = \lim_{t \to \infty} \eta(\sigma^t(q^0)) \quad ; \quad \text{uniformly in } t \geq 0 \tag{4.4.8}$$

$$= \eta(q^*) \quad ; \quad \text{by Theorem 4.3.1, the contraction property}$$

and since convergence in (4.4.8) is uniform in t, $\lim_{t \to \infty}$ and $\lim_{N \in \Gamma_*}$ commute.

5. FINAL REMARKS AND FURTHER CONVERGENCE RESULTS

Theorems 3.1 and 3.2 show that at any finite time t or at equilibrium, with probability tending to one as N tends to infinity, the system process can be found in an arbitraily small neighborhood of the approximating process. However, it does not immediately follow that performance statistics (such as those presented in the companion paper [4]) derived from the system process converge to the corresponding estimates derived from the approximating process. A question that remains, for example, is whether the equilibrium utilization of any one memory node converges to the estimate $\lambda^* = p(\alpha - (q^*)'(1))$, given by the root $\lambda = \lambda^*$ in $[0,1]$ of the quadratic equation

$$\alpha = \frac{\lambda}{p} + \frac{\lambda^2}{2(1-\lambda)} .$$

To conclude, we show in this Section that all moments of the queue length at a memory node converge to corresponding estimates derived from the approximating process. It follows, in particular, that the equilibrium memory utilization converges to λ^*.

For $q(z) = \sum_{i \geq 0} q_i \, z^i$, define

$$m_k(q) = \sum_{i \geq 0} q_i \, i^k .$$

With respect to the system process $(q^t)_{t \geq 0}$, $m_k(q^t)$ is a random variable, which may be regarded as the empirical k-th moment of the queue length at time t. By symmetry, its expectation $Em_k(q^t)$ is the k-th moment of the queue length distribution at any given memory node at time t.

The following Corollary of Theorem 3.1 follows from the facts that (i) $m_k(q)$ is a uniformly continuous function from any compact subset of Ω_α to the reals, and (ii) (μ_N^t) is a tight family of probability measures, by Theorem 4.2.4.

Let q^∞ denote a random state over $\Omega_{N,M}$ with the equilibrium measure π_N.

Corollary 5.1: Suppose that for some $A>0$, the initial state $q^0(z)\in K_A$ for every $N\in\Gamma_\alpha$. Then, for any $\beta > 0$,

$$\lim_{N\in\Gamma_\alpha} \Pr\{|m_k(q^t) - m_k(\sigma^t(q^0))| > \beta\} = 0 \text{ ; uniformly for } t>0$$

$$\lim_{N\in\Gamma_\alpha} \Pr\{|m_k(q^\infty) - m_k(q^*)| > \beta\} = 0 .$$

The difficulty in proving the counterpart for Em_k is that m_k is not bounded, so the weak convergence result of Theorem 3.1 does not in itself imply consistent convergence of Em_k. To see what might go wrong, consider the sequence of probability measures $(\xi_n)_{n\geq 1}$ on the nonnegative integers where $\xi_n(0)=1-1/n, \xi_n(n)=1/n, \xi_n(i)=0$ for i not equal to 0 or n. Though $\xi_n\Rightarrow\xi_*$ where $\xi_*(0)=1, \xi_*(i)=0$ for $i>0$, the sequence of expectations $\sum i\xi_n(i)=n$ does not converge to $\sum i\xi_*(i)=0$. This difficulty for Em_k is overcome via the following Lemma.

Suppose $(\xi_n)_{n\geq 0}$ is a tight family of probability measures. We say that a continuous function η is tight with respect to $(\xi_n)_{n\geq 0}$ if the family of measures $(|\eta|\xi_n)_{n\geq 0}$ is tight; that is, there is a sequence of compact sets $(K_m)_{m\geq 0}$ such that

$$\lim_{m\to\infty} \int_{K_m} |\eta| d\xi_n - \int |\eta| d\xi_n = 0 \text{ uniformly for } n \geq 0 .$$

By adapting the argument used in the proof of Theorem 4.2, it can be shown that any function η satisfying

$$\lim_{A\to\infty} \limsup_{q(R)>A} |\eta(q)| / q(R) = 0 ,$$

for fixed $A>0$ is tight with respect to the family of measures (μ_N^t). Also, for every $k \geq 1$, the functions $m_k(q)$ are tight with respect to (μ_N^t).

The following result shows that if a sequence of probability measures converges weakly then, for any tight function, the corresponding sequence of expectations of that function converges consistently. The proof is an application of Prohorov's theorem [2] and is omitted.

Lemma 5.2: If the continuous function η is tight with respect to $(\xi_n)_{n\geq 0}$ and ξ_n weakly converges to ξ^*, then $\lim_{n\to\infty} \int \eta d\xi_n = \int \eta d\xi^*$.

Combining Lemma 5.2 with Theorem 3.1 gives:

Corollary 5.3: Suppose that for some $A > 0$, the initial state $q^0(z) \in K_A$ for every $N \in \Gamma_\alpha$. Then

$$\lim_{N \in \Gamma_\alpha} (\mathrm{E}[m_k(q^t)] - m_k(\sigma^t(q^0))) = 0 \ ; \quad \text{uniformly for } t \geq 0 \ .$$

$$\lim_{N \in \Gamma_\alpha} (\mathrm{E}[m_k(q^\infty)] - m_k(q^*)) = 0 \ .$$

Last, we note the equilibrium utilization of any one memory node is given by

$$\lambda_N = p(N/M - \mathrm{E}m_1(q^\infty)) \ .$$

The approximating process provides the corresponding estimate: $p(\alpha - (q^*)'(1))$. Corollary 5.3 shows that this estimate is asymptotically exact.

6. APPENDIX: PROOF OF THE CONTRACTION PROPERTY

In this section, we provide the proof of Theorem 4.3.1, which is restated here for convenience.

Theorem 4.3.1: For some $R > 1$, any $1 < \rho < R$, and any $A > 0$,

$$\lim_{t \to \infty} \| \sigma^t(q) - q^* \|_\rho = 0 \quad ; \quad \text{uniformly for } q \in K_A \ ,$$

where q^* is the unique fixed point of σ.

Before proving Theorem 4.3.1, we establish a few lemmas. First, let us extend the definition of stochastic domination, $\leq_{st}$, to generating functions. Let Ω denote the space of all probability generating functions with radius of convergence no less than R. Thus, $\Omega_\alpha \subset \Omega$. For all r, s in Ω_α, define

$$r \leq_{st} s \text{ if and only if } \sum_{i=0}^{n} r_i \geq \sum_{i=0}^{n} s_i \text{ for all } n \geq 0 \ .$$

Note that if $r \leq_{st} s$, then $r_0 \geq s_0$ and $r'(1) \leq s'(1)$ and furthermore, $r(\rho) \leq s(\rho)$ for all $\rho \in (1,R)$. Define $\lambda(q) = p(\alpha - q'(1))$ and, for any $\lambda \geq 0$, define the operator S_λ by

$$S_\lambda q(z) = \frac{q(z)e^{\lambda(z-1)} - q(0)e^{-\lambda}}{z} + q(0)e^{-\lambda} \ .$$

Thus, $\sigma(q) = S_{\lambda(q)}\, q$. Let q_λ denote the fixed point of operator S_λ in Ω, $S_\lambda q_\lambda = q_\lambda$ and

$$q_\lambda(z) = \frac{(1-\lambda)(z-1)}{1 - e^{\lambda(z-1)}} \ .$$

Note that the fixed point q^* coincides with q_{λ^*}, where λ^* is such that $p(q_{\lambda^*})'(1) + \lambda^* = \alpha p$. Hence, $\lambda^* = \lambda(q^*)$.

The following propositions are straightforward to prove:

$$\begin{array}{llll} \forall \lambda > 0 & r,s \in \Omega & r \leq_{st} s \Rightarrow S_\lambda r \leq_{st} S_\lambda s \\ \forall \lambda_1, \lambda_2 > 0 & \forall q \in \Omega & \lambda_1 \leq \lambda_2 \Rightarrow S_{\lambda_1} q \leq_{st} S_{\lambda_2} q \end{array}$$

For any subset X of Ω, define

$$\bar{X} = \{ q \in \Omega : \exists x \in X \ , \ q \leq_{st} x \} \ .$$

Conversely, define

$$\underline{X} = \{q \in \Omega : \exists x \in X , x \leq_{st} q\} .$$

Finally, for any $\varepsilon > 0$ and q in Ω, let $B_\varepsilon(q)$ denote the ball with center q and radius ε, with respect to our metric $\|\cdot\|_\rho$.

The start of our investigation is a result lifted from the proof of Lemma 4.2.1 in Section 4.2.

Lemma A1: For all $A > 0$, all $q \in K_A$, there is a positive integer a and $0 < \delta < 1$, such that

$$\lambda(\sigma^a(q)) < 1 - \delta .$$

Henceforth, when we refer to δ below we mean the δ satisfying Lemma A1. The proof of the following simple result is omitted.

Lemma A2: For all $q \in \Omega_x$ and all $\varepsilon > 0$ there exists $\beta > 0$ such that $\overline{B_\beta(q)} \cap \underline{B_\beta(q)} \subset B_\varepsilon(q)$

The next lemma is a by-product of the upper bound analysis (Section 4.2).

Lemma A3: For all $A > 0$, there exists $B > 0$ such that for all $n \in N$ and all $q \in K_A$ we find $\sigma^n q \in K_B$.

Proof: We know, by Lemma A1, that there exists an integer a and a real number $\delta > 0$ such that $\forall q \in \Omega_\alpha : \lambda(\sigma^a q) < 1 - \delta$. Since Ω_α is closed with mapping σ, we have $\forall n \in N$, $\forall q \in \Omega_\alpha : \lambda(\sigma^{n+a} q) < 1 - \delta$. Therefore $\forall q \in \Omega_\alpha \; \forall n \in N : \sigma^{a+n} q \leq_{st} S^n_{1-\delta}\, \sigma^a q$, which implies, for any $R > 1$, $\sigma^{a+n} q(R) \leq S^n_{1-\delta} \sigma^a q(R)$. For all q in Ω we have $S_{1-\delta} q(R) \leq \dfrac{e^{(1-\delta)(R-1)}}{R}\, q(R) + 1$. Having chosen $R > 1$ to satisfy Lemma 4.2.3 of Section 4.2, we see R satisfies $\dfrac{e^{(1-\delta)(R-1)}}{R} < 1$. Finally, $\forall n \in N, \forall q \in \Omega$,

$$S^n_{1-\delta}\, q(R) \leq \left[\frac{e^{(1-\delta)(R-1)}}{R}\right]^n q(R) + \frac{1}{1 - \dfrac{e^{(1-\delta)(R-1)}}{R}} \qquad \blacksquare$$

Lemma A4: For any $\lambda \leq 1 - \delta$ and any compact subset $K \subset \Omega$,

$$\lim_{n \to \infty} S^n_\lambda\, q = q_\lambda$$

uniformly for all $q \in K$.

Proof: This is a classical result for the M/D/1 queue. The result holds for all metrics $\|\cdot\|_\rho$ with $\rho \le R$ and $\frac{e^{\lambda(R-1)}}{R} < 1$. ■

Now suppose the quantity $A > 0$ is fixed. Define $\bar{\lambda}_n$ and $\underline{\lambda}_n$ by

$$\bar{\lambda}_n = \max_{q \in K_A, t \ge 0} \lambda(\sigma^{n+t}q)$$

$$\underline{\lambda}_n = \min_{q \in K_A, t \ge 0} \lambda(\sigma^{n+t}q) \ .$$

Note that $\bar{\lambda}_n < \infty$, by Lemma A3 and the continuity of the functions $\lambda(q)$ in K_B. The sequence $\bar{\lambda}_n$ is decreasing to the limit $\bar{\lambda}$. The sequence $\underline{\lambda}_n$ is increasing to the limit $\underline{\lambda}$. Obviously $\underline{\lambda} \le \bar{\lambda}$.

Lemma A5: For every $\varepsilon>0$ there is a positive integer N such that for all $n \ge N$, the set $\sigma^n(K_A) \subset \overline{B_\varepsilon(q_{\bar{\lambda}})} \cap \underline{B_\varepsilon(q_{\underline{\lambda}})}$.

Proof: Since q_λ is a continuous function of λ on Ω, there exists $\beta > 0$ such that $B_\beta(q_{\bar{\lambda}+\beta}) \subset B_\varepsilon(q_{\bar{\lambda}})$. Having fixed β there exists $N' \ge 0$ such that, $\forall n \ge N$, $\bar{\lambda}_n \le \bar{\lambda} + \beta$. Therefore $\forall n \ge 0, \forall q \in K_A$,

$$\sigma^{n+N}q \in \overline{S^n_{\bar{\lambda}+\beta}(\sigma^N(K_A))} \subset \overline{S^n_{\bar{\lambda}+\beta}(K_B)} \ .$$

By Lemma A4, there is a positive integer N'' such that, $\forall n \ge N''$,

$$S^n_{\bar{\lambda}+\beta}(K_B) \subset B_\beta(q_{\bar{\lambda}+\beta}) \ .$$

Thus, $\forall n \ge N' + N''$,

$$\sigma^n(K_A) \subset \overline{B_\beta(q_{\bar{\lambda}+\beta})} \subset \overline{B_\varepsilon(q_{\bar{\lambda}})} \ .$$

The proof of $\sigma^n(K_A) \subset \underline{B_\varepsilon(q_{\underline{\lambda}})}$ is similar, and relies on the fact that, $\forall N, n \ge 0$, we have $\sigma^{n+N}(K_A) \subset \underline{S^n_{\underline{\lambda}_N}(K_B)}$ ■

Lemma A6: If $\bar{\lambda} = \underline{\lambda} = \lambda^*$ then Theorem 4.3.1 holds.

Proof: Lemma A5 shows that $\forall \varepsilon > 0$ there exists $N \ge 0$ such that, $\forall n \ge N$, $\sigma^n(K_A) \subset \overline{B_\varepsilon(q_{\bar{\lambda}})} \cap \underline{B_\varepsilon(q_{\underline{\lambda}})}$. The result then follows from Lemma A2 and $\bar{\lambda} = \underline{\lambda} = \lambda^*$. ■

Thus, the proof of Theorem 4.3.1 reduces to showing that $\bar{\lambda} \leq \lambda^*$ and $\underline{\lambda} \geq \lambda^*$. The proofs of these two inequalities rest on the following Lemma.

Lemma A7: For any compact subset $K \subset \Omega_\alpha$ and $1-\delta > \lambda > \lambda^*$, there is an integer b such that, for all $q \in K$, there is some $k \leq b$ satisfying $\lambda(\sigma^k q) < \lambda$.

Proof: Let $K_N = \{q \in K: \forall n \leq N, \lambda(\sigma^n q) \geq \lambda\}$. Each K_N is compact. Suppose that $\bigcap_{N \geq 0} K_N \neq \emptyset$, and choose q such that, $\forall n \geq 0$, we have $\lambda(\sigma^n q) \geq \lambda$. Therefore, $\forall n \geq 0$, $S_\lambda^n q \leq_{st} \sigma^n q$. We know $\lim_{n \to \infty} S_\lambda^n q = q_\lambda$ since $\lambda \leq 1 - \delta$. In addition, $(q_\lambda)'(1) > (q^*)'(1)$, so there exists n satisfying

$$(\sigma^n q)'(1) \geq (S_\lambda^n q)'(1) > (q^*)'(1) .$$

Since $\lambda(\sigma^n q) = p(\alpha - (\sigma^n q)'(1))$ it follows that $\lambda(\sigma^n q) < p(\alpha - (q^*(1)) = \lambda^*$, which contradicts $\lambda > \lambda^*$. Thus, we know $\bigcap_{N \geq 0} K_N = \emptyset$. Moreover, $\bigcup_{N \geq 0} C_{K_N} \supset K$ where C_X denotes the complement of X in Ω (i.e., $X \cup C_X = \Omega, X \cap C_X = \emptyset$). Since the C_{K_N} are open and increasing ($C_{K_N} \subset C_{K_{N+1}}$), there exists a finite b such that $K \subset C_{K_b}$, thus $K_b = \emptyset$. ■

Lemma A8: For any compact subset $K \subset \Omega_\alpha$ and $\lambda < \lambda^*$, there is an integer b such that, for all $q \in K$, there is some $k \leq b$ satisfying $\lambda(\sigma^k q) > \lambda$.

Proof: The proof is the same as the proof of the previous lemma, reversing each inequality except those between integers.

Corollary A9: For any $\lambda > \lambda^*$, there is an integer b such that, $\forall n \in N$ and all $q \in K_A$, ther is some integer k between 0 and b such that $\lambda(\sigma^{n+k} q) < \lambda$.

Corollary A10: For any $\lambda < \lambda^*$, there exists an integer b such that $\forall n \in N$ and for all $q \in K_A$ there exists an integer k between 0 and b such that $\lambda(\sigma^{n+k} q) > \lambda$.

Proof: For all $n \in N$, all $q \in K_A$, we know $\sigma^n q \in K_B$. ■

Lemma A11: For $\lambda > 0$ and all $q \in \{\overline{q_\lambda}\} \cap \Omega_\alpha$ we have

$$\lambda(\sigma q) = \lambda \text{ and } \lambda(q) \leq \lambda \Rightarrow \lambda(q) = \lambda .$$

Proof: We know

$$\lambda(\sigma q) = (1 - p)\lambda(q) + p(1 - e^{-\lambda(q)} q(0)) .$$

Since $q \in \{q_\lambda\}$ and $\lambda(q) \leq \lambda$, $1 - e^{-\lambda(q)} q(0) \leq 1 - e^{-\lambda} q(0) \leq 1 - e^{-\lambda} q_\lambda(0) = \lambda$, so $\lambda(\sigma q) \leq \lambda$. Equality holds only if $\lambda(q) = \lambda$ and $q(0) = q_\lambda(0)$. ■

Lemma A12: For $\lambda > 0$ and all $q \in \{q_\lambda\} \cap \Omega_\alpha$ we have

$$\lambda(\sigma q) = \lambda \text{ and } \lambda(q) \geq \bar{\lambda} \Rightarrow \lambda(q) = \lambda .$$

Proof: The proof is similar to the previous one. ■

Lemma A13: For all $\varepsilon_1 > 0$, there exists $\varepsilon_2 > 0$ and $N \geq 0$ such that for all $q \in K_A$ and all $n \geq N$ the following points hold

(i) $\lambda(\sigma^n q) < \bar{\lambda} - \varepsilon_1 \Rightarrow \lambda(\sigma^{n+1} q) < \bar{\lambda} - \varepsilon_2$

(ii) $\lambda(\sigma^n q) > \underline{\lambda} + \varepsilon_1 \Rightarrow \lambda(\sigma^{n+1} q) > \underline{\lambda} + \varepsilon_2$.

Proof: Suppose point (i) be not true. Then there exists a sequence N_n of integers tending to infinity and a sequence $q^n \in \sigma^{N_n}(K_A)$ such that $\lambda(q^n) < \bar{\lambda} - \varepsilon_1$ and $\liminf_{n \to \infty} \lambda(\sigma q^n) \geq \bar{\lambda}$; this last condition, by the definition of $\bar{\lambda}$, reduces to $\lim \lambda(\sigma q^n) = \bar{\lambda}$. The sequence of q^n is entirely in K_B and is thus relatively compact. Let q be one of its limit points. We have, of course, $\lambda(q) \leq \bar{\lambda} - \varepsilon_1$ and $\lambda(\sigma q) = \bar{\lambda}$. We also have $q \in K_B$. Since $q^n \in \sigma^{N_n}(K_A)$ and $N_n \to \infty$, Lemma A5 implies $q \in \{q_{\bar{\lambda}}\}$. The facts $q \in \{q_{\bar{\lambda}}\}$, $\lambda(q) < \bar{\lambda}$ and $\lambda(\sigma q) = \bar{\lambda}$ contradict Lemma A11, so point (i) is proven. The proof of point (ii) is similar, but uses Lemma A12. ■

Lemma A14: $\bar{\lambda} = \underline{\lambda} = \lambda^*$.

Proof: Let us suppose $\bar{\lambda} > \lambda^*$. As a simple consequence of Lemma A13 we have that for all $\varepsilon_1 > 0$ and for all integer b, there exists $N \geq 0$ and $\varepsilon_2 > 0$ such that, $\forall q \in K_A$ and $\forall n \geq N$,

$$\lambda(\sigma^n q) < \bar{\lambda} - \varepsilon_1 \Rightarrow \forall i = 1,2,\ldots,b : \lambda(\sigma^{n+i}) < \bar{\lambda} - \varepsilon_2 .$$

Using Corollary A9, identifying $\bar{\lambda} - \varepsilon_1$ with λ, we know that $\forall n \geq 0$ and $\forall q \in K_A$ there exists k between 0 and b such that $\lambda(\sigma^{n+k} q) < \bar{\lambda} - \varepsilon_1$. Thus, with $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$, $\forall n \geq N + b$,

$\forall q \in K_A$, $\lambda(\sigma^n q) < \bar{\lambda} - \varepsilon$, which contradicts the definition of $\bar{\lambda}$. Therefore $\bar{\lambda} \leq \lambda^*$. The proof that $\underline{\lambda} \geq \lambda^*$ is similar. ■

REFERENCES

[1] Baskett, F. and Smith, A.J., "Interference in multiprocessor computer systems with interleaved memory", *Communications of the ACM*, 19, 6, pp. 327-334 (July 1976).

[2] Billingsley, P., *Convergence of Probability Measures*, John Wiley & Sons, 1968.

[3] Billingsley, P., *Probability and Measure*, second edition, John Wiley & Sons, 1986.

[4] Boguslavsky, L.B., Greenberg, A.G., Jacquet, P., Kruskal, C.P., and Stolyar, A.L., "Simple Models of Memory Interference in Multiprocessors, Part I: Approximate Analysis", submitted.

[5] Greenberg, A.G., Jacquet, P., Kruskal, C.P., "Analyis of Memory Conflicts", presented at the ORSA/TIMS meeting on Analysis and Control of Large Stochastic Systems, Chapell Hill, North Carolina, USA, (May 1988), unpublished.

[6] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input versus output queueing on a space-division packet switch", *IEEE Transactions on Communications*, COM-35, 12 (Dec. 1987), 1347-1356.

[7] B. R. Rau, "Interleaved memory bandwidth in a model of a multiprocessor computer system", *IEEE Transactions on Computers*, C-28, 9 (Sept. 1979),

[8] C. Skinner and J. Asher, "Effect on storage contention on system performance", *IBM System Journal*, vol. 8, 319-333, 1969.

[9] B. Smilauer, "General model for memory interference in multiprocessors and mean value analysis", *IEEE Transactions on Computers*, C-34, 8 (Aug. 1985), 744-751.

[10] Stolyar, A.L., "Asymptotics of stationary distribution for one class of closed service networks." Preprint: Moscow Institute for Problems of Data Transmission, R.L. Dobrushin (Ed.), Moscow, 1988 (in Russian).

[11] W. D. Strecker, "An analysis of the instruction execution rate in certain computer structures", Ph.D. dissertation, Carnegie-Mellon Univ., June 1970.

[12] Yen, W.C. and Fu, K.S., "Performance Analysis on Multiprocessor Memory Organizations", ACM Pacific '80, Distributed processing, new directions for a new decade, San Francisco (November 1980), 142-153.

[13] Yen, D.W.L., Patel, J.H., and Davidson, E.S., "Memory interference in synchronous multiprocessor systems." *IEEE Transactions on Computers*, C-31, 11 (Nov. 1982), 1116-1121.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249 - 6399